Universidad de Alcalá

Programa de Doctorado en Tecnologías de la Información y las Comunicaciones

Effective Neuro-Evolutionary Schemes for Solar Radiation Estimation Problems

Tesis Doctoral presentada por

ADRIÁN AYBAR RUIZ

2021

Universidad de Alcalá

Programa de Doctorado en Tecnologías de la Información y las Comunicaciones

Effective Neuro-Evolutionary Schemes for Solar Radiation Estimation Problems

Tesis Doctoral presentada por

ADRIÁN AYBAR RUIZ

Directores:
Dr. SANCHO SALCEDO SANZ
Dra. SILVIA JIMÉNEZ FERNÁNDEZ

Alcalá de Henares, 2021

Abstract

This Ph.D. thesis' goal is focused on the optimization of renewable energy resources development, specifically solar PV energy, using different hybrid computational Machine Learning techniques.

Energy is the engine of our society, allowing us performing almost every action taken by human beings in our daily routine, and providing a constant evolution and development in all our fields. Currently, fossil fuels entail the higher percentage of energy sources in our planet. They have several advantages, such as easy and constant production, but, at the same time, they present substantial disadvantages, like the extreme pollution associated with these resources, and their contribution to global warming and climate change. This is the reason why the largest and most powerful economies are working for a energy change towards renewable sources for a sustainable development. In the introduction of this thesis, a large number of studies are presented, which foresee a penetration by over a 50% of this kind of energies in the next decades. Strong investments are being made in this field, looking for technology development and, besides, introducing these energies into society as a matter to be taken into account, for it is related to economic and social status. However, the development of energy systems mainly based on renewable energy will surely be slow, since these energies depend on variables which are out of our control, mainly atmospheric and climatic variables, which are intrinsically intermittent. This matter must be taken into account, due to the amount of energy demanded by society at the present, as well as the tremendous increase that is predicted for this demand in the future, due to new technologies and new ways of daily routines, like electric vehicles or IoT, etc.

To obtain a solution for this problem, on one hand, it is fundamental to achieve the capability of predicting the quantity of energy obtained in each moment, avoiding increases or decreases of energy, being this matter the core of this Ph.D. Thesis, where the optimal feature selection for predicting the quantity of global solar radiation in a given point is studied. On the other hand, all the information to do the prediction process will be obtained from a numerical weather mesoscale model called WRF (Weather Research and Forecasting), a static model based on different physic equations which involve different variables like humidity, pressure or percentage of cloud fraction in any point and different heights in the planet. Additionally, dynamic information, like global solar radiation can be obtained from a radiometric measuring point in Toledo, Spain, allowing us to get a database of the solar global radiation in the past few years. The result of

mixing both of these data will be added as inputs in our hybrid systems.

In this work, a deep analysis in the state of art for machine learning models is performed, so as to solve the problems previously considered. Different contributions have been proposed:

1. One of the pillars of this work is focused on the optimal feature selection in the exploitation of solar PV radiation in a given point. For this purpose, Extreme-Learning Machine (ELM) will be used as regressor element in the system, where the output of the ELM will be calculated from the WRF outcome features added as inputs in the system.

2. The second contribution of this thesis is related to parameters selection problems. More specifically, the use of EAs such as Grouping Genetic Algorithm (GGA) or Coral Reef Optimization (CRO) hybridized with others ML are used as classifiers and regressors. Regarding this, the GGA or CRO look for several subsets of basic parameters to solve the problem, and the regressor employed provides the prediction in terms of the selected by the Genetic Algorithm (GA), reducing the computational cost maintaining a good accuracy.

Finally, the several of the mentioned algorithms are applied in the same problem already defined, in order to get the global solar radiation prediction in different points, dealing to improve previous results in other works and obtaining new applications and techniques, as new paths of research in the future.

Resumen en Castellano

Esta Tesis doctoral tiene como objetivo la optimización de la explotación de recursos energéticos renovables, siendo en este caso el objetivo principal la energía solar fotovoltáica y la predicción del recurso. El análisis llevado a cabo se fundamenta en la utilización de algoritmos y técnicas computacionales meta-heurísticas híbridas, especf́icamente algoritmos genéticos de agrupamiento, algoritmos de arrecife de coral y redes neuronales artificiales.

La energía es el motor de nuestra sociedad, que nos permite realizar casi la totalidad de las acciones que el ser humano hace a diario, propugnando una constante evolución y desarrollo en cualquier ámbito de nuestra vida. Actualmente, los combustibles denominados fósiles son la mayor fuente energética de nuestro planeta, tanto por la facilidad y continuidad en su obtención, como porque gran parte de la tecnologá hasta hace poco tiempo estaba basada en su consumo. Es sobradamente conocido que este tipo de energía son un recurso finito y extremadamente contaminante, afectando al medio en el que vivimos. Debido a esto, las grandes economías están apostando por un cambio paulatino en las fuentes de energía, siendo los recursos renovables los que garantizan un futuro sostenible para nuestra forma de vida. Como veremos en la introducción de esta tesis, numerosos estudios preveen una penetración de dichas fuentes por encima del 50% en las próximas décadas. Las fuertes inversiones que se realizan en esta materia, para su desarrollo y ampliación en la penetración, hacen de ellas un elemento interesante a niveles socio-económicos. Esta evolución será lenta y paulatina, ya que este tipo de energías tienen un elemento que no podemos controlar (al contrario que en las energías fósiles) que es su intermitencia intrínseca, ya que son energías que dependen de en muchos casos de fenómenos naturales, fundamentalmente atmosféricos en el caso de la energía solar y eólica. El objetivo final es conseguir que la penetración de estas energías renovables supere a las energías fósiles y permita salvar el problema anteriormente descrito.

La mejor opción para conseguir una penetración elevada de los recursos renovables es desarrollar sistemas de predicción que permitan establecer el recurso disponible en el futuro próximo, para poder adelantarnos a cualquier exceso o defecto de la producción. Este será un elemento clave sobre el que gira la presente Tesis doctoral. Especf́icamente, la selección de las variables más importantes para realizar una predicción del recurso solarcon el menor error posible y que nos permita realizar las gestiones oportunas en tiempo y forma.

En este aspecto, obtendremos toda la información para la predicción de un modelo numérico-

meteorológico de meso-escala (WRF), el cual es un modelo estático y que, basado en distintas ecuaciones físicas, aporta una serie de variables como la humedad, la presión o el porcentaje de cielo nublado que obtendremos en cualquier lugar del planeta, a distintas alturas. Este modelo puede ser utilizado para la predicción en sí mismo, pero usualmente sufre de falta de resolución en la estimación de la radiación solar en un punto específico de la superficie. En esta Tesis se utiliza una red neuronal de entrenamiento rápido para realizar la predicción final en el punto deseado, usando los algoritmos metaheurísticos mencionados anteriormente para la selección de las variables óptimas a usar en esta predicción.

En este trabajo se ha realizado un análisis del estado del arte de los modelos de aprendizaje máquina que se utilizan actualmente, con el objetivo de resolver los problemas asociados a los temas tratados con anterioridad. Diferentes contribuciones han sido propuestas:

1. Uno de los pilares esenciales de este trabajo está centrado en la selección de los parámetros más importantes en la medición de la energía solar fotovoltaica en un punto dado. Con este propósito, el algoritmo de Extreme Learning Machine (ELM), será utilizado como elemento regresor del sistema, que permitirá calcular en un tiempo de cómputo mínimo, el valor de la radiación solar global en función de las variables de entrada.

2. Otro de los aspectos tratados está relacionado con problemas de selección de características para una óptima predicción de nuestro sistema regresor. Concretamente, con el uso de algoritmos evolutivos como Algoritmos de Agrupación Genética (GGA) o los algoritmos de Optimización de Arrecife de Coral (CRO) hibridados con otros métodos de aprendizaje máquina como clasificadores y regresores. En este sentido, el GGA o CRO analizan diferentes conjuntos de características para obtener aquel que resuelva el problema con el menor posible, y el regresor empleado proporciona la predicción en funcioÌn de las características obtenidas por el GGA, reduciendo el coste computacional con gran fiabilidad en los resultados.

Los diferentes algoritmos mencionados han sido aplicados a una serie de casos reales con medidas de radiación solar del observatorio Astronómico de Toledo, España.

Agradecimientos

Que difícil es agradecer en tan pocas palabras cuando estás rodeado de tantas grandísimas personas que he tenido la suerte de tener tan cerca en todo este tiempo, tanto las que vienen impuestas, como las que eliges a lo largo de tu vida. Creo en la casualidad y también en la causalidad, ya que, nos definimos a lo largo de nuestra vida por nuestras acciones y nuestros sueños, y de ahí, la afinidad con el resto de personas que se cruzan en nuestro camino. Y ya sea por "causa" o por "casualidad", considero que soy una persona afortunada.

Empecemos por mis directores de tesis, Sancho y Silvia, Silvia y Sancho, no solo personas de las que he podido absorber conocimiento técnico, sino de vida, humano, las cuales siempre me han ayudado, me han entendido y me han consentido, en definitiva, grandes amigos que me han enriquecido como persona y como profesional día a día. Continuo probablemente con la persona que más ha influenciado mi camino en la vida, junto con mis padres, Antonio, "Porti", confió en mí cuando llegué a una entrevista fuera de plazo, en chandal y en la que solo hablamos del mundo, y no del objeto en cuestión. Muchos años después, aprendí de él que todos somos iguales, que cursamos caminos diversos en función de nuestras capacidades, esfuerzos y objetivos, pero en definitiva, iguales. Me inculcó el espíritu emprendedor, institucional, social, me ayudó a entender que cada momento es una oportunidad y que cada experiencia es un crecimiento, y no solo lo inculca, si no que lo práctica día a día, siempre diré de vosotros tres que sois los talentos más increíbles que me he cruzado en mi camino y que sois dueños de vuestra vida, ya que elegisteis el camino de la educación y gracias a ello, he sido uno de los grandes afortunados de disfrutar esa elección. Sé que debería ser al revés, pero qué orgulloso estoy de vosotros, y de la suerte que he tenido al haberme permitido aprender de vosotros todos estos valores que a día de hoy me ayudan en mis proyectos, tanto profesionales, como de vida.

Gracias a su vez, al resto del grupo, Enrique, Lucas y a otros compañeros que han estado en dicho camino, como Javier del Ser o Carlos Casanova. Agradezco la oportunidad de haber trabajado con vosotros en mayor o menor medida, la ayuda que me habéis brindado y como facilitáis nuestra vida en este difícil camino que elegimos.

Llegan los "tiparracos", mis compañeros, y ojo... qué compañeros! Laura y Carlos, Carlos y Laura, qué dos cabezas, qué dos grandes personas, qué dos grandísimos amigos. Leales, inteligentes, divertidos, cariñosos... qué feliz he sido y cómo os echo a día de hoy de menos. Sé que tenéis grandes carreras por delante, como era de esperar, pero no sabéis cuanto me alegro.

Siempre estuvisteis y siempre estaréis ahí, nos veremos con poca frecuencia pero no por ello deja de ser especial. Ahora disfrutáis enseñando todo lo que nos enseñaron a nosotros, vuestras universidades tienen que dar gracias cada día de los valores humanos que tienen y ojalá, algún día, en algún Máster o similar, tenga la suerte de que seáis mis profesores y poder disfrutar de vosotros, que seguro así será.

La familia, esa que no se elige y que en función de donde caigas, tienes gran parte de tu vida, personalidad y posibilidades definidos a priori. Yo que puedo decir más que la suerte que he tenido, que he caído de pie y que mi familia es una familia completamente genial, trabajadora, con valores, buenas personas y que quieren lo mejor tanto para ellos como para los demás. Ana y Alfonso han creído siempre en mí, me han permitido decidir, equivocarme, aprender, luchar, me han enseñado el coste de cada una de las cosas que tiene la vida, el esfuerzo como herramienta principal del éxito y que a día de hoy, pues han permitido que sea quien soy, con mis puntos positivos y negativos, con mis éxitos y mis fracasos, pero en definitiva, siempre me han dejado decidir mientras ellos estaban ahí. Qué importante es dar ejemplo, qué importante es ver como tu padre trabaja cada día hasta la extenuación y jamás quejarse o que tu madre busque mil maneras para que todo esté bien aunque todo este mal, porque al final, todo eso queda dentro de uno mismo, esas enseñanzas, el valor de tenerlo todo y no tener nada. Muchísimas gracias por todos y cada uno de los regalos que me habéis brindado, ¡Qué fácil me habéis puesto el camino!. Ahora alguien que siempre ha sido importante, pero por los defectos de la juventud uno es incapaz de verlo hasta que llega a una madurez que te quita la neblina, o las tonterias según quiera llamarlo, de la cabeza. Alicia, ¡Qué hermana tengo!, qué joven y qué inteligente, qué capacidad de superarse, de tener las cosas claras, de saber lo que uno quiere, que grandísima persona. En los últimos tiempos, donde más he podido disfrutar de ti y que no voy a dejar de hacer jamás, me he quedado prendado de lo que eres, de lo que transmites y de lo que generas a tu alrededor, cambiaría todas y cada una de mis virtudes por tener solo una de las tuyas. Sé que tu camino será especial y que siempre decidirás bien, si es que lo increíble de todo esto es, ¡Qué se supone que yo soy el mayor! El resultado de esto es parte tuyo, con esas horas y horas de charla, en las que me cuidabas cuando no estaba en mis mejores momentos. En fin, has conseguido algo que creo que es lo mejor que te puedo ofrecer, mi respeto, ser la persona que más respeto en el mundo y dotar de valor todas y cada una de tus opiniones. Te quiero enana.

Hace años, una persona muy importante en mi vida se fue, aunque no quisiera, hablaba de Porti como una de las mayores influencias a nivel “profesional/intelectual”, en este caso es la mayor inspiración que tendré, mi Yaya, Isabel. Primero quiero decirte te echo muchísimo de menos y ojalá estuvieras aquí para disfrutarlo conmigo, esto te haría más ilusión que a cualquiera. Recuerdo cuando te decía “Yaya, tienes que aguantarme eh, que quiero que me veas”, y me contestabas "¡Ay hijo! ójalá, pero ya estoy muy vieja". Siempre creíste en mi y me decías, “no te preocupes si todo te va a salir bien, que tu eres muy listo”. Lo recuerdo y me río. Listo no sé si seré, pero gitano Yaya... ¡un rato! Como me enseñaste a luchar, aprendiste sola a leer, manejabas la familia, el dinero y nunca faltaba de nada, conseguiste todo lo que te propusiste, fuera fácil o difícil, sufrías tú para que no sufrieran los demás, todo lo que hiciera falta. Si la vida fuera justa, hubieras vivido una vida de película por puro merecimiento. Solo quiero decirte que me acompañas y en todas y cada una de mis decisiones importantes, te imagino con esos ojitos pequeños detrás de las gafas, mirándome fijamente ayudándome a decidir, “si todo te va a salir bien”. No sé si será el cerebro que es muy práctico o que eras tan espectacular como recuerdo, pero es que todos y cada uno de los recuerdos que tengo de ti, son perfectos, inmejorables. Todo lo que he hecho y todo lo que haré, siempre serán para ti Yaya.

A mis demás abuelos, a mis tíos, ¡Juancar como me metiste la vena ingeniera!, Silvi cómo te he querido siempre, has sido una hermana mayor más que una tía, mis primos, Iris como compañera de batallas (ese peque que viene), ¡gracias por estar ahí! Y para terminar, pues los amigos, los que estuvieron, los que están y lo que estarán, Martita, como me ayudó en los peores momentos y me animaba cuando ya no quería seguir a dar un pasito más, como me enseñaste a currar y trabajar duro, cuando estudiabas horas y horas para sacar tus exámenes, como a tantas otras cosas, eres una grandísima persona, leal, trabajadora y por ello, estoy seguro que todo te irá bien, porque tienes claro que lo importante es rodearte de gente buena y que lo importante es ser buena con los demás. Samu, tiaco, eres el amigo perfecto, siempre siempre siempre estás ahí, que claro lo tienes todo, solo necesitas leer un mensaje u oírme cinco segundos para saber si necesito ayuda o que debes estar ahí, ojalá fuera como tú. Soraya, mi rubia, nos vemos poco pero nos queremos lo mismo cada día, será imposible cambiarlo. Los viejos como yo les llamo, en especial a álvaro por todo lo que me has cuidado y enseñado, como te has preocupado por mi y como siempre has creído apoyarme en lo que era lo mejor para mi, me siento como tu hermano pequeño, ¡gracias!, a Juanjo y Mario, que son mis dos pepitos grillos, Juanjo más bicho, una persona con la experiencia y la categoría como tú, pero la juventud y la energía de un chaval, ¡te echo de menos!. Mario es la voz de la sabiduría, gracias por enseñarme y orientarme en los pasos correctos y por intentar que siempre tome decisiones buenas y correctas. Pablete, “Jiu”, nos conocimos hace relativamente poco, pero eres de esas personas que entra como una apisonadora en tu vida y que nunca te falla, que con solo mirarte sabe que piensas, te conozca de un mes o de una vida, serás un grandísimo amigo siempre, eres un crack. Sergio, por entrenarme y apostar por mi, por escucharme como si fueras un psicólogo más que un entrenador, no sé si me has ayudado o me has entrenado más. Iñaki y Alex, que dos personajes, que trio maléfico generamos, que ilusión me hace veros de tanto en cuanto. Navas, siempre serás una estrella del rock, talento y desorden, eres mágico, como mola que trabajemos juntos en Ari. Noita, como me cuidas, vas y vienes, pero cuando ves que algo falla, siempre estás ahí, no olvidaré jamás el primer mensaje que recibí tuyo, solo por eso, nunca dejaré de cuidarte. Nuri, ¡qué difícil eres!, te propusiste coger a un chaval y convertirle en un tío hecho y derecho, y vaya que si lo has conseguido, como siempre dices “le vamos puliendo poco a poco, pero aún queda camino”, cuánta razón llevas, gracias por todo lo que me has enseñado. Miguelillo, titán, como tu te haces llamar a veces, "Dios", que grandes amigos y compañeros de curro somos, como es posible tener una mano derecha que sabes que nunca te falla, que sueña contigo y que lucha por tus mismos sueños, te digo algo, tarde o temprano lo conseguiremos y llegaremos al objetivo de alcanzar a millones de personas, recuérdalo. Laurita, Clasing eres tú, le das sentido a todo, sin ti probablemente

Clasing ni existiría, que profesional, que responsable, que buena persona, ¡qué haría yo sin ti!, ¡ná de ná!. Ojalá y siempre vayamos juntos los tres locos y sigamos siendo más y más en la family con nuestros chicos de Clasing, Pablete, Leti y los que vendrán. Y para terminar, Laura Lph, como así apareces desde que te conocí en mi teléfono, recuerda que siempre serás tú la inteligente de los dos aunque no lo reconozca a menudo, gracias por ayudarme a terminar esto dándome caña, preguntándome, agobiándote por los dos y haciéndome ver que incluso aunque no haga falta, siempre se puede hacer más aunque sea por puro hobby, sigue luchando porque lo tienes todo, pero todo, para ser lo que quieras en la vida, eso si, disfruta del camino, del esfuerzo y sobretodo, no dejes de ser como eres, tienes un gran futuro y lo mejor es que te lo mereces, que nadie te ha regalado nada, me sorprendiste el primer día viendo el lanzamiento del Falcon 9 y unos meses después solo puedo admirarte. No sé que deparará la vida, pero que no vas a dejar de luchar por lo que quieres, lo tengo claro, menuda cabezota tienes. Además recordarte, que si a día de hoy puedo escribir estas líneas es en gran parte gracias a ti, a que ese 2 de Julio me llamaras para tomar una cerveza, recuerda que te debo una y que espero no devolvértela jamás.

Me dejaré gente, seguro, os pido disculpas, sabéis que pienso en todos y que si me necesitáis, estaré. Sí que os digo que gracias a todos vosotros, a día de hoy, soy lo que soy. Por todos vosotros, puedo decir que soy una persona afortunada.

A. Aybar

La mayoría de corredores no corren porque quieran vivir más. Lo hacen porque quieren vivir al máximo.

Haruki Murakami

Contents

List of Figures

List of Tables

LIST OF ACRONYMS

AEMET Meteorological State Agency of Spain

AFWA Air Force Weather Agency

AI Artificial Intelligent

ANFIS Adaptive Neuro-Fuzzy Inference SYstem

ARMA Autoregressive-Moving-Average

CRO Coral Reef Optimization

CRO-SL Coral Reef Optimization with Substrate Layer

CRO-SP Coral Reef Optimization with Species

DE Differential Evolution

DNI Direct Normal Irradiance

ELM Extreme-Learning Machine

EU European Union

FAA Federal Aviation Administration

FSL Forecast Systems Laboratory

FSP Feature Selection Problem

GA Genetic Algorithm

GGA Grouping Genetic Algorithm

GHI Global Horizontal Irradiance

GSR Global Solar Radiation

GW Gigawatt

HS Harmony Search

IEA International Energy Agency

LSTM Long-Short-Term Memory

ML Machine Learning

MLP Multi-Layer Perceptron

MO-CRO Multi-Objective Coral Reef Optimization

MODIS Moderate Resolution Imaging Spectroradiometer

NCAR National Center for Atmospheric Research

NCEP National Centers for Environmental Prediction

NN Neural Network

NWM Numerical Weather Models

PBL Planetary Boundary Level

PSO-ELM Particle Swarm Optimization and Extreme-Learning Machine

PV Photo Voltaic

RBF Radial Basis Functions

RBR Red Blue Ratio

RMSE Root Mean Squared Error

RRTM Rapid Radiative Transfer Model

SDS Sustainable Development Scenario

SGD Stochastic Gradient Descent

SME Small and Medium-sized Enterprises

SVM Support Vector Machine

SVR Support Vector Regression

US United States

USA United States of America

WEKA Waikato Environment for Knowledge Analysis

WMO World Meteorological Organization

WRF Weather Research and Forecasting meso-scale model

WSM Weather Solar Model

YSU Yonsei University

Part I

Ph.D. Thesis by Compendium of Publications

Publications of the compendium

This Ph.D. Thesis is composed of three main articles which form the core of the research work carried out. The first two articles summarize the work carried out in the development of Grouping Genetic Algorithms (GGA). The first one is a development of a hybrid GGA approach, in this case with specific application to a problem of telecommunication network design, but with extension to other optimization problems. In fact, the second article of the compendium describes the application of the GGAs to a problem of solar radiation prediction from numerical weather model, where the groups of the GGA deal with a problem of Feature Selection. The last article of the compendium presents a different approach to the problem, in which a Coral Reefs Optimization with Species is proposed in a problem of solar radiation prediction.

1. L. Cuadra, **A. Aybar-Ruiz**, M. A. del Arco, J. Navío-Marco, J. A. Portilla-Figueras and S. Salcedo-Sanz, "A Lamarckian Hybrid Grouping Genetic Algorithm with repair heuristics for resource assignment in WCDMA networks," *Applied Soft Computing*, vol. 43, pp. 619-632, 2016. (JCR: 3.541, Q1)

2. **A. Aybar-Ruiz**, A., S. Jiménez-Fernández, L. Cornejo-Bueno, C. Casanova, J. Sanz-Justo, P. Salvador-González and S. Salcedo-Sanz, "A novel Grouping Genetic Algorithmâ€"Extreme Learning Machine approach for global solar radiation prediction from numerical weather models inputs," *Solar Energy*, vol. 132. pp. 129-142, 2016. (JCR: 4.018, Q1)

3. S. Salcedo-Sanz, S. Jiménez-Fernández, **A. Aybar-Ruiz**, C. Casanova-Mateo, J. Sanz-Justo and R. García-Herrera, "A CRO-species optimization scheme for robust global solar radiation statistical downscaling," *Renewable Energy*, vol. 111, pp. 63-76, 2017. (JCR: 4.357, Q1)

Part II

Part I: Introduction and Computational Methods

Chapter 1

Introduction

Today's world is completely dependent on electricity, mainly produced based on fossil fuels such as petroleum, natural gas or coal. Moreover, according to [112], the world primary energy demand is projected to expand by almost 60% from 2002 to 2030, with an average increase of 1.7% per year. However, the high environmental impact of current energy resources, together with the need for addressing the impact of climate change, led to an important development of renewable energy sources [76]. The International Energy Agency (IEA) has stated that electricity generation from renewable energy is expected to rise up to 39% by 2050 [13]. In fact, renewable sources have experienced a huge growth in the last two decades, and they are today thought as the resources which will erase fossil fuels from our society in the next fifty years. The prevalence of renewable resources has several challenges, that must be solved before renewable energies overtake fossils as primal energy resources. The most important issue with renewable energy sources is to manage their inherent intermittence, which currently avoid that renewable energies overpass 40% of penetration in the energetic mix.

Among renewable resources, solar, wind, geothermal, biomass and hydroelectric energy sources are the main types of renewable energy. However, solar resource is currently thought as the future renewable source, due to the extraordinary solar resource we have available. More specifically, solar energy is a clean, extremely abundant and sustainable source of energy [37], that poses a low risk to the environment. Solar energy development has been specially important in mid-east and Southern Europe countries, plus North America, where the solar resource is able to be exploited all year around [56], because the percentage of clear sky and the cloud non-appearance is higher than in the rest of the world. Among these renewable resources, solar photovoltaic was observed as the one of the world leading renewable energy sources.

1.1 Solar energy production in the world

Solar PV generation overtook bioenergy and is now the third-largest renewable electricity technology in absolute figures, after hydropower and onshore wind. Moreover, Solar PV generation increased 22% (+131 TWh) in 2019, representing the second-largest absolute generation growth of all renewable technologies, slightly behind wind and ahead of hydropower. Despite decelerating growth due to recent policy changes and uncertainties in China (the largest solar PV market globally), 2019 was a year of record global growth in PV capacity. As competitiveness continues to improve, solar PV is still on track to reach the levels envisioned in the SDS, which will require average annual growth of 15% between 2019 and 2030.

Solar PV electricity generation increased by 131 TWh globally in 2019, to research 720 TWh, second only to wind in absolute terms, to account for 2.7% of the electricity supply. This growth was significantly lower than in 2018, however, because global solar PV capacity additions stalled in 2018 and China's deployment further contracted in 2019. This was mainly as a result of a sudden change in China's solar PV incentives to curb costs and address grid integration challenges to achieve a more sustainable PV expansion. The European Union, India and the United States contributed equally to the solar output increase. Solar PV generation rose sharply in Southeast Asia, driven by a surge in new capacity in Vietnam from 0.1 GW to 5.4 GW. Capacity additions increased in the United States, the European Union, Latin America, the Middle East and Africa, which together compensated for the slowdown in China, resulting in a record year for PV deployment. Solar PV is well on track to reach the Sustainable Development Scenario (SDS) level by 2030, which will require electricity generation from solar PV to increase 15% annually, from 720 TWh in 2019 to almost 3300 TWh in 2030.

Stimulated by strong policy support concentrated mostly in Europe, the United States and Japan, deployment of distributed solar PV systems in homes, commercial buildings and industry has been growing exponentially over the last decade. In most countries, commercial and residential systems already have electricity generation costs that are lower than the variable portion of retail electricity prices. The increasing economic attractiveness of distributed PV systems could therefore lead to a rapid expansion in the coming decades, attracting hundreds of millions of private investors.

Regarding the solar energy growth by countries, in Germany, solar photovoltaic energy was promoted in early 1970s, and now is a leading country in PV development in Europe [51]. In 2019, German PV plants produced about 46.5 TWh, an increase of 1.7 percent compared to 2018. In Japan, a lot of technology-push policies were promulgated for solar energy utilization sine 1950s [129]. In 2019 the annual share of solar PV power generation in Japan increased from 6.5% in the previous year to 7.4%.

Comparing with other developed countries, the photovoltaic energy industry started relatively late in China, but it experienced an explosive increase since 2000s [137]. Solar PV capacity in China slowed for the second year in row to 30.1 GW in 2019. This expansion is significantly lower than the 53.1 GW in 2017, when the government phased out feed-in tariffs and introduced deployment quotas (in June 2018) to control costs and tackle grid integration challenges. Overall, this policy shift is expected to make solar PV technology more cost-competitive within and outside China, leading to more sustainable development over the longer term. A large number of subsidy-free projects were already in development in multiple chinese provinces in 2019. Distributed solar PV capacity is expected to increase rapidly in China, driven by new auctions for commercial and industrial applications and subsidies for residential systems.

Australia is a leading country in solar PV, with the highest level of the annual global solar radiation [97]. In spite of this, solar energy only accounts for only 3% of the total power consumed in the country, which remains behind wind energy (5%) and hydro-power (7%).

Growth in the United States was stable, with 13.2 GW of solar PV becoming operational in 2019, one quarter higher than 2018 additions, as a result of federal tax incentives and state-level policies. In the European Union, solar PV additions increased 98% year-on-year in 2019 owing to faster deployment in Spain, Germany and the Netherlands. Brazil installed a record-level 2.1 GW of new solar PV capacity in 2019, more than doubling its achievement in 2018. Generous net metering incentives stimulated this rapid expansion, as residential and small commercial consumers receive significant returns on their investments. US and EU growth increased and higher Latin American capacity addition expansion was led by Brazil.

Covid-19 has led to construction delays, supply chain disruptions and weaker investment in the PV sector in 2020. Utility-scale projects are susceptible to supply chain concerns, labour constraints and construction delays, all leading to delays in project commissioning. The distributed PV sector is more at risk as it relies on both individuals and SMEs, who are more severely affected by lockdown measures and any economic downturn resulting from Covid-19. Despite the slowdown expected in 2020, acceleration of PV capacity deployment is likely to continue in the medium-term, as the cost of electricity generation from solar PV is increasingly cheaper than alternatives. The rapid recovery of the distributed PV sector will depend on the pace of economic recovery and government policies. The impact of the Covid-19 crisis on PV deployment is extensively covered in the *IEA Renewable Energy Market Update* released in May 2020 [85].

1.2 Solar radiation prediction and estimation problems

As the majority of renewable resources, solar power production is intrinsically stochastic, and significant variations in solar energy production occur due to the presence of clouds, atmospheric dust or particles [80]. This intermittency can cause problems to include solar production in the energetic mix. The best option to manage this intermittence of the solar resource is to predict the solar production of the most important plans, in the same way that wind production is predicted in wind farms.

The basis to estimate solar radiation at any given location is to apply the classical astronomical equations [109]. In addition, the well-known different processes that modifies solar radiation when it passes through the atmosphere can also be considered for a better approach: molecular (or Rayleigh) scattering by the permanent gases, aerosol (or Mie) scattering due to particles and the abortion by different gases. Nevertheless, this scheme is very simple and more reliable and robust methods are necessary to deliver accurate predictions. In this sense, many different approaches have been proposed to estimate or predict solar radiation amount in the last years.

For longer prediction time horizons (e.g. 6 hours or more), physics-based or numerical models are usually employed [28, 22, 49]. In the last years, these numerical models have been improved by the use of Machine Learning (ML) or Artificial Intelligence (AI) algorithms for solar radiation prediction or estimation problems, as we will further detail in Chapter 2. The problem of solar energy prediction usually involves the accurate prediction of the solar radiation at a given point (the solar plant facility), and this prediction depends completely on different atmospheric variables [92, 50, 57, 114, 119]. The majority of these approaches include different inputs based on geographical and atmospheric parameters such as latitude, longitude, temperature, wind speed and direction, sunshine duration or precipitation [69, 74], among others.

1.3 Motivation and objective of the work

In the last years, there have been a massive interest in new algorithms for solar radiation prediction, as will be detailed in Chapter 2. Many of the most successful approaches are in fact ML-based approaches, which have exploited the capacity of ML algorithms of extracting information from data. Some of these ML algorithms have just applied some of the state-of-the-art algorithms such as different types of Neural Networks (NN), Support Vector Regression (SVR), Random Forest, etc. to a set of databases, with very little knowledge of the physics behind the problem. This may lead to poor results in some cases, mainly when the prediction time horizon is larger than 6 hours. In this cases, it is of major importance to take into account the atmospheric conditions in order to obtain a robust prediction system for solar radiation.

There are basically two ways of including atmospheric information in ML prediction problems: one is to directly hybridize ML algorithms with numerical methods which provide the atmospheric future state. In general it is not an easy task, and major problems of compatibility, computational performance, etc. may arise. The second possibility is to consider input variables which directly considers the atmospheric state, both as direct measurements whenever possible, or variables from numerical models, such as meso-scale models or re-analysis etc. In [98] it is possible to find a review of the most important data sources for Earth observation problems, including data for prediction problems in renewable energy. This second case is usually the most used, since avoid the problems of hybridization and computational complexity of a direct hybridization with numerical models. However, there are other issues when considering this second possibility. The most important one is that the number of predictive (input) variables can be huge depending on the specific prediction problem, which in turn, may lead to poor performance of ML approaches. Thus, it is completely necessary to apply Feature Selection mechanisms [96] to improve the performance of ML algorithms in solar radiation and estimation problems.

In this Ph.D. Thesis, we propose a novel family of hybrid approaches based on ML for solar radiation prediction and estimation problems. The hybrid meta-heuristic proposed include a feature selection mechanism based on evolutionary algorithms, and a fast training neural network (Extreme Learning Machine (ELM) [44]) which is able to obtain accurate results within a very small computation time. Specifically, in a first proposal, we evaluate a novel grouping genetic algorithm (GGA) scheme as feature selection mechanism. The proposal consists of evaluating the input features in each group formed in the GGA, and keeping the best performance as final feature set for the ELM. Different modifications of this scheme including alternative crossover and evolution dynamics are proposed and evaluated in this Thesis, for solar radiation prediction problems. A second algorithmic proposal is based on a new version of the Coral Reefs Optimization (CRO) algorithm [93, 94], an recently proposed evolutionary approach which has been successfully applied to different optimization problems. In this case we exploit an advanced version of the CRO in which different species are considered within the algorithm, each representing a possible encoding for an optimization problem. Each encoding represents in this case a given number of inputs for the ELM in the solar radiation prediction problem. This way, the Coral Reefs Optimization with Species (CRO-SP) is able to evolve different encodings, looking for the best one within a single population of the algorithm. In this case the CRO-SP is also hybridized with a ELM to carry out the solar radiation prediction.

With these ideas in mind, the main objective of this Ph.D. Thesis is to evaluate the performance of these new family of hybrid ML algorithms in specific problems of solar radiation prediction, considering real data and input variables that consider the atmospheric state, both direct measurements and data from numerical weather models.

1.4 Structure of the Thesis

The rest of this Thesis has been structured in the following way: Chapter 2 reviews the most important previous work on ML algorithms for solar radiation prediction, and also the work in the last years about Feature Selection, with a specific review of Feature Selection methods in solar radiation prediction problems. Chapter 3 is devoted to describe the most important algorithms and computational methods which will be used to construct the proposed family of hybrid algorithms for solar radiation prediction. Specifically, we will review the most important characteristics of Grouping Genetic Algorithms, Coral Reefs Optimization algorithms and Extreme Learning Machines. The experimental part of the Thesis will be presented in Chapters 4 and 5, in which different problems of solar radiation prediction and estimation are discussed. Chapter 4 presents a novel hybrid approach formed by a Grouping Genetic Algorithm for feature selection and an Extreme Learning Machine as predictor. In turn, chapter 5 presents the performance of a hybrid Coral Reefs Optimization algorithm with Species, also hybridized with an Extreme Learning Machine as predictive approach. The Thesis is closed with Chapter 6, where the most important conclusions of this work are summarized. Finally, Chapter 6.2 presents some future lines of research that this work has opened.

Chapter 2

Previous work: ML approaches in solar prediction and algorithms for feature selection problems

2.1 Machine learning for solar radiation prediction problems

The first models able to estimate the expected solar radiation with acceptable accuracy were developed in the twentieth century. Probably the most well-known empirical model was the one suggested by [5]. This model that estimates global solar radiation amount from sunshine duration data, was subsequently modified by taking into account some other relevant meteorological variables [109]. Despite their simplicity, the most serious constraint of this type of models is that they are based on the assumption that the time series data used to predict the incoming solar radiation are linear. However, the atmospheric system is chaotic and highly non-linear, which inevitably affects the predictability of these models. To overcome this limitation, several authors proposed the use of Numerical Weather prediction Models (NWM). These systems are able to model the dynamics of the atmosphere, as well as the physical processes involved, by employing a set of equations based on physical laws of motion and thermodynamics. Therefore, they can be considered as a reliable approach to estimate incoming solar radiation. For example, in [80] an interesting review on the performance of different NWP models to forecast solar irradiance in the US, Canada and Europe is presented. The evaluation includes forecasts based on global, multiscale and mesoscale NWP models. Additionally, NWP models have demonstrated their ability to provide useful solar radiation climate data sets, like the work of [79] where a 60 years (1950 to 2010) climatology of incident shortwave downward solar radiation at the surface over the Iberian Peninsula is obtained from simulations performed using the WRF model. An alternative approach is presented in [66], where an artificial neural network ensemble model is proposed for the prediction of global solar radiation by exploiting information within infrared Meteosat channels.

In the last years, many different approaches have been proposed for global solar radiation prediction, a lot of them using Machine Learning or Computational Intelligence techniques. The majority of these approaches include different inputs based on geographical and atmospheric parameters such as latitude, longitude, temperature, wind speed and direction, sunshine duration, precipitation, etc. [69, 74]. According to [12], sunshine duration, air temperature and relative humidity are the most widely used meteorological parameters to predict daily solar radiation and its components. All these parameters are well correlated with the daily solar global radiation as pointed out in [122]. In [67] a Bayesian framework for artificial neural networks, named as automatic relevance determination method, was developed to evaluate the more relevant input parameters in modelling solar radiation. In fact, neural computation paradigm has been massively applied to this prediction problem, like in [8], where it is shown that Radial Basis Functions (RBF) neural networks obtain excellent performance in the estimation of solar radiation. In [30] a comparison between Multi-Layer Perceptrons (MLP) and RBF neural networks in a problem of solar radiation estimation is carried out. Experiments in eight stations in Oman show the good results obtained with the neural algorithms. A similar approach, also comparing MLPs and RBFs (with different predictive variables) has been recently proposed in [7], in this case the authors test the neural network with data obtained in Iran. In [78] the performance of a MLP in a problem of solar radiation prediction in time series is compared to that of ARMA, Bayesian inference, Markov Chains and k-Nearest Neighbors models, for specific problems in Corsica and Southern France. Another work dealing with solar radiation time series prediction is [133], where a hybrid algorithm that involves an ARMA model and a time-delay neural network is proposed. In [43] a neural network to predict hourly solar radiation in a region of Turkey is proposed. The paper also introduces a 2D model for solar radiation useful for visualization and data inspection. In [35], the forecasting of solar irradiance proposed utilizes features extracted from all-sky images, such as the number of cloud pixels, frame difference, gradient magnitude, intensity level, accumulated intensity along the vertical line of sun or the number of corners in the image. Other works on solar radiation prediction involve ARMA models, as [53] a hybrid approach based on ARMA and time delay neural networks has been successfully tested in data from a solar station in Singapore. Another paper involving hybrid ARMA and neural networks is [120], where this hybrid approach is successfully applied to solar radiation prediction in different cities of the French Mediterranean coast and Corsica. Alternative approaches that apply neural networks as prediction methodology also include novel predictive variables, such as satellite data [110] or temperature and relative humidity [84]. Other machine learning algorithms, such as Support Vector Regression (SVR) algorithms have been also applied to solar radiation prediction problems from meteorological predictive variables [21, 136]. Specifically, a least-square SVM is proposed in that work, comparing the results obtained with that of auto-regressive and RBF neural networks. In [82] the potential of multi-layer perceptron neural networks with back-propagation training algorithm is shown in a problem of global solar radiation estimation in Iran. Results comparing the performance of the neural networks with that of an empirical equation for global solar radiation prediction (Hargreaves and Samani equation) show good performance of the neural approach. In [11] a hybrid approach that includes hidden Markov models and generalized fuzzy models has been proposed and tested in real solar irradiation data in India. Finally, we discuss very recent hybrid approaches proposed to problems of solar energy prediction, such as [75] where a SVR has been hybridized with a fire-fly algorithm to select the best parameters of the SVR, or [70], where a hybrid SVR–Wavelets approach is presented in a problem of horizontal global solar radiation prediction. The goodness of this novel approach has been tested in a real problem of solar radiation estimation in Bandar Abbas (Iran). Moreover, in [28], a post-processing technique (Kalman filtering) is used to improve the hour-ahead forecasted Global Horizontal Irradiance (GHI) from 1) the measured GHI at the ground, and 2) the Weather Research and Forecasting (WRF) meso-scale model, and results at Reunion Island are provided.

In more recent years (from 2017 to 2020), the use of Machine Learning-based (ML) algorithms [38] has consolidated as a reliable alternative/complement to NWP in solar energy prediction problems. For example, [63] proposed a new day-ahead spatiotemporal prediction method for solar radiation in order to ensure the efficient operation of power systems. In [128] solar and wind energy resources in North Korea were studied by using both satellite data and NWP reanalysis variables. [39] proposed a deep learning hybrid model to predict solar radiation in two phases, firstly a convolutional network is used to extract features, and secondly, a Long-Short-Term Memory (LSTM) network is applied for the prediction stage. Moreover, [81] also used LSTM for hourly day-ahead solar irradiance prediction achieving competitive results for a case of study in Santiago, Cape Verde. [47] explored the blending of four models (Satellite, WRF-Solar, Smart Persistence and CIADCast) by using Support Vector Machines aiming to improve GHI and DNI forecasts in the Iberian Peninsula; [3] studied several categories of ML techniques such as artificial neural networks, gradient boosting trees or classification and regression trees, among others, for solar radiation estimation at two different locations, Turkey and USA; [27] applied ELMs to forecast long-term incident solar radiation over Australia using relevant satellite-based input data extracted from the moderate resolution imaging spectroradiometer and enriched by geo-temporal input variables; [33] evaluated the use of a hybrid Particle Swarm Optimization and ELM (PSO-ELM), to accurately predict daily global solar radiation on seven stations located on the Loess Plateau of China during 1961-2016; [73] evaluated the performance of a hybrid neural network with simulated annealing in a problem of solar radiation prediction in Iran; [26] proposed a novel pipeline combining artificial neural networks with satellite-derived (MODIS) land surface temperature (LST) for forecasting long-term global solar radiation; [48] suggested a hybrid model based on random forest and a firefly algorithm for a problem of global solar radiation prediction in Malaysia. More recently, [64] assessed the application of deep learning algorithms as a solar predictor, using a new strategy based on the portfolio theory in order to reduce predictability errors. Furthermore, [86] undertook regression analysis of solar irradiance in Iran, paying attention to the effect of different types of storage tanks and of changes in longitude and latitude. [77] proposed a recurrent neural network model to investigate how emerging deep learning algorithms contribute to accurate solar radiation prediction, specially in comparison to swallow artificial neural networks. Some state-of-the-art reviews of ML models widely used in solar energy prediction problems can be found in the works [121, 4, 72].

As it has been mentioned, some approaches consider the use of satellite images together with ML techniques to estimate incoming solar radiation. For example, [135] uses fuzzy logic and neural networks for estimating hourly global radiation from satellite images. In [87] an ELM combined with satellite data and geographic variables is applied to a solar radiation prediction problem over Turkey. [118] assessed the progress made by two different sources of reanalysis data, ERA5 and COSMO-REA6, in the estimation of surface irradiance. [134] proposes a deep learning method based on embedding clustering and functional deep belief networks to estimate solar radiation with data from a total of 30 stations from China. More recently, [23] evaluates the performance of several ML regression techniques (multi-layer perceptrons, ELMs, Support Vector Regressors and Gaussian Processes) in a problem of global solar radiation estimation from geostationary satellite data.

Different approaches discussing Extreme Learning Machine (ELM, a novel training method for artificial neural networks) applications in solar radiation prediction problems have been recently proposed, such as [87], where the ELM approach is applied to a solar radiation prediction problem from satellite measures. In [2] a case study of solar radiation prediction in Saudi Arabia is discussed comparing the performance of artificial neural networks with classical training and ELMs. In [29] a hybrid wavelet-ELM approach is tested in a problem of solar radiation prediction for application in a photovoltaic power station. Finally, in [91] a comparison of a support vector regression algorithm and an ELM is carried out in a problem of direct solar radiation prediction, with application in solar thermal energy systems, and in [92], where a hybrid ELM–Coral Reefs Optimization is proposed for solar radiation prediction in Southern Spain. Finally, in [97] developed a model by combining CRO with ELMs, where CRO works as a feature selection function guided by the ELM algorithm, to predict daily global solar radiation in the Sunshine State of Australia, achieving a competitive performance. That work is fully related to this Ph.D. Thesis development.

2.2 Feature selection

2.2.1 Feature Selection in Machine Learning

Feature selection is an important task in Machine Learning related problems because irrelevant features, used as part of the training procedure of different prediction systems, can increase the cost and running time of the system, and make its generalization performance much poorer[15, 131]. In its more general form, the FSP for a learning problem from data can be defined as follows: given a set of labeled data samples $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_l, y_l)$, where $\mathbf{x}_i \in \mathbb{R}^n$ and $y_i \in \mathbb{R}$ (or $y_i \in \{\pm 1\}$ in the case of classification problems), choose a subset of m features ($m < n$), that achieves the lowest error in the prediction of the variable y_i.

There are two main paradigms/general approaches to face a FSP:

- The *wrapper approach* to the FSP was introduced in [55]. In this approach, the feature selection algorithm conducts a search for a good subset of features using the classifier/regressor itself as part of the evaluating function. Figure 2.1(a) shows the idea behind the wrapper approach: the classifier/regression technique is run on the training dataset with different subsets of features. The features subset which produces the lowest estimated error in an independent but representative test set is chosen as the final feature set. For further reading on wrappers methods, the following classical works can be consulted [62, 126, 88]. In the case of the wrapper method, the FSP admits a mathematical definition as follows: The FSP consists of finding the optimum n-column vector $\boldsymbol{\sigma}$, where $\sigma_i \in \{1, 0\}$, that defines the subset of selected features, which is found as

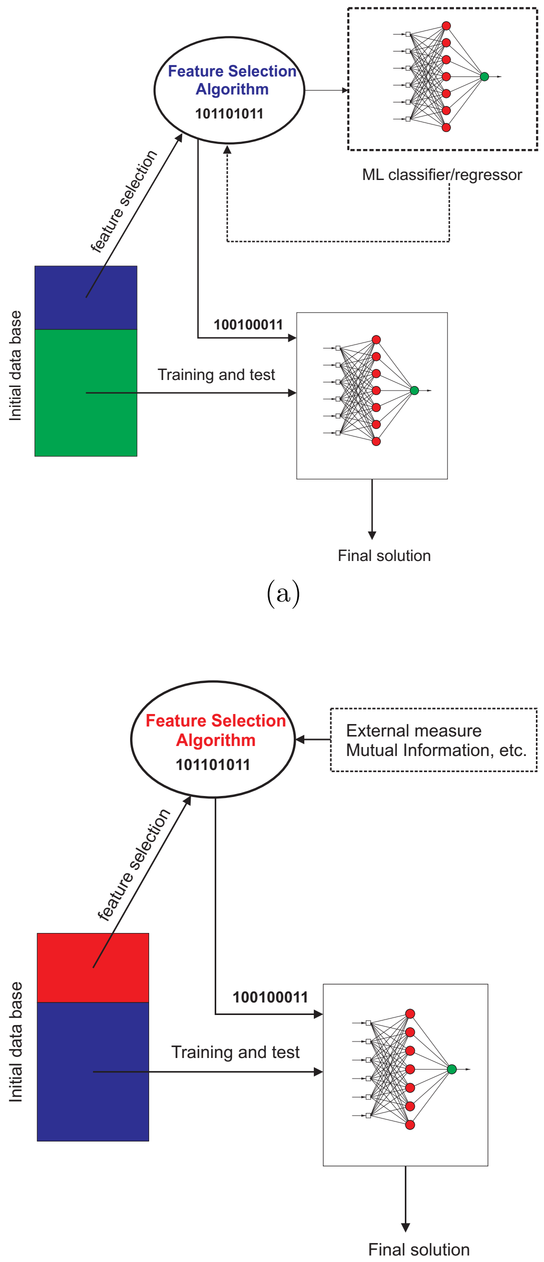
$$\boldsymbol{\sigma}^o = arg \min_{\boldsymbol{\sigma},\boldsymbol{\alpha}} \left(\int V(y, f(\mathbf{x} * \boldsymbol{\sigma}, \boldsymbol{\alpha})) dP(\mathbf{x}, y) \right), \tag{2.1}$$

  where $V(\cdot, \cdot)$ is a loss functional, $P(\mathbf{x}, y)$ is the unknown probability function the data was sampled from and we have defined $\mathbf{x} * \boldsymbol{\sigma} = (x_1\sigma_1, \ldots, x_n\sigma_n)$. The function $y = f(\mathbf{x}, \boldsymbol{\alpha})$ is the classification/regression engine that is evaluated for each subset selection, $\boldsymbol{\sigma}$, and for each set of its hyper-parameters, $\boldsymbol{\alpha}$.

- In the *filter approach* to the FSP, the feature selection is performed based on the data, ignoring the classifier algorithm. An external measure calculated from the data must be defined in order to select a subset of features. After the search, the best feature subset found is evaluated on the data by means of the classifier algorithm. Note that filter algorithms performance completely depends on the measure selected for comparing feature subsets. Figure 2.1(b) shows an example of how a filter algorithm works. Filter methods are usually faster than wrapper methods. However, their main drawback is that they totally ignore the effect of the selected feature subset on the performance of the classification/regression algorithm during the search, so usually their performance is poorer compared to wrapper approaches. Further analysis and application of filter methods can be found in [115, 116].

- There are finally different works which have combined both wrapper and filter methodologies to come up with hybrid approaches, which have been reported to have a very good performance in specific applications [113, 45, 46, 34].

For both wrapper and filter methods, a binary representation can be used for the FSP, where a 1 in the i_{th} position of the binary vector means that the feature i is considered within the subset of features, and a 0 in the j_{th} position of the binary vector means that feature j is not considered within the subset of features. Note that using this notation is equivalent to encode the problem as the vector $\boldsymbol{\sigma}$ included in expression (2.1). Note also that there are 2^n different subsets of features (being n the total number of features), and the problem is to select the best one in terms of a certain measure, which can be either internal (wrapper methods) or external (filter methods) to the classifier.



(a)

(b)

Figure 2.1: (a) Outline of a Wrapper method; (b) Outline of a Filter method.

2.2.2 Feature selection in solar energy

In [35] a system for solar irradiance very short-term prediction (minutes time-horizon) is proposed. The work uses a solar irradiance prediction scheme with features extracted from all-sky images. The idea is to obtain proper features from all-sky images derived from an all-sky camera sited in Taiwan, apply a feature extraction algorithm to the images, and then use a regression technique to predict a clearness index from them. In a second step, the clearness index is used to calculate the desired solar irradiance together with the extraterrestrial solar irradiance value, which only depends on astronomical variables. The features considered for the clearness index estimation are related to the all-sky images obtained: number of cloud pixels, where a RBR threshold method is applied to decide cloud pixels or not. Frame difference between images in t and $t-1$ moments, which is important to locate moving clouds. Gradient magnitude, which provides information corresponding to edges in an image. In this case, edges are often related to cloud boundaries in all-sky images. Intensity level, since in the all-sky images the brightness level of clear skies and cloudy skies are very different. Accumulated intensity along the vertical sun line. The idea of this feature is that if the sunlight is strong, the line caused by the sun would cross the entire image, whereas in cloudy situations, the vertical line of sun would not cross the image. Number of corners, since corners in all-sky images correspond to the details of the clouds that have edges with at least two directions in a local patch of the sky. Using these features, the proposed system applies a filter feature selection to different images consisting in ranking the features which have a higher correlation to the clearness index. The most relevant features are then used in a linear regression system to estimate the clearness index from the image. The results in real data from all-sky images in Taiwan collected in August 2011 show that the proposed system is an accurate tool to estimate short-term solar irradiation prediction locally. The experimental results shows an improvement in the short-term prediction of solar irradiance of about 4% in comparison with the estimation of the solar irradiance directly from weather variables.

In [124] a study of the main influencing input parameters for solar radiation prediction with neural networks is carried out in different locations of India, by using the Waikato Environment for Knowledge Analysis (WEKA) software. Different variables such as daily average temperature, minimum temperature, maximum temperature, altitude, sunshine hours and site location are considered. The study uses a previous wrapper method with a regression tree implemented in WEKA to select the best set of features. After this process, several neural networks from WEKA are evaluated on this best set of features. Improvements over 13% are obtained after the FSP process (comparing the neural networks without feature selection pre-processing) in some of the locations considered. Note that also in this case, the small number of features involved in the problems made possible an exhaustive search algorithm, which was implemented in the WEKA software.

In [130] a feature extraction method is applied to select the best set of input parameters for a Support Vector Machine (SVM) classifier, in order to reconstruct a database of weather types. These weather types are directly connected with the photovoltaic power generation accuracy, so the most useful features are extracted from photovoltaic data, and they serve as inputs in a SVM, to reconstruct the database of weather types. As can be seen, this is a classification problem used as a previous step in the estimation of photovoltaic power generation prediction for buildings.

In [83] a problem of forecasting the electricity power generation by a solar photo-voltaic system is tackled. Short-term prediction (from 5 to 60 minutes ahead) is considered. Input variables at different previous times are considered: solar irradiance, temperature, humidity and wind speed. A total of 4200 variables are finally available as inputs to predict the photo-voltaic generation in the next hour, in intervals of 5 minutes. A filter method is used to select the best set of variables for the prediction, in this case the correlation-based feature selection, which selects the best set of variables with higher correlation with the objective variable. After this feature selection, two machine learning algorithms are applied to generate the system's prediction: an ensemble of neural networks and a support vector regression approach. Experiments in power data collected from the St. Lucia campus of the University of Queensland in Brisbane, Australia, have shown that the neural network ensemble is able to outperform the SVR, obtaining solutions around 1% better after applying the filter feature selection method proposed.

In [71] an adaptive neuro-fuzzy inference system (ANFIS) has been applied to select the most influential variables in a daily horizontal diffuse solar radiation prediction problem. Relevant variables are considered in order to study how different groups predict solar radiation: daily diffuse and global solar radiation on a horizontal surface, sunshine duration, minimum air temperature, maximum air temperature, average air temperature, relative humidity, water vapor pressure, daily maximum possible sunshine duration, solar declination angle and extraterrestrial solar radiation on a horizontal surface. Four years of measured data from the Iranian Meteorological Organization (January 2009 to December 2012) have been used. Analysis of the best combination of 3 variables have been carried out, showing the %age of improvement when considering 1, 2 or 3 variables. The best result obtained (3 variables) was over 30% better than the solution of the system considering just 1 variable. No comparison results with alternative approaches were shown in the work, which makes difficult a further analysis of the proposal's performance.

In [132] a hybrid niching genetic algorithm – linear regression approach is used to estimate global solar radiation in El Colmenar, Argentina. Data from 14 different weather stations were used in this work. Climatic variables such as daily average temperature, air humidity, atmospheric pressure, cloudiness, and sunshine hours are considered. The idea is to reconstruct the global solar radiation at El Colmenar from the data in the other 13 measurement stations. A niching genetic algorithm with binary encoding is considered, in which a 1 stands for including a given variable in the prediction, and a 0 stands for not including it. The linear regression is used due to its good computational complexity performance and its interpretability, though the authors admit that neural networks can obtain better results. The prediction obtained is only compared with the niching genetic algorithm with different number of individuals and generations, with the best prediction result obtained with 200 individuals and 150 generations, improving a 3% the solution of the algorithm with 50 individuals and 35 generations. A complete comparison with alternative approaches is not provided.

Finally, [123] offers a first general review of some works dealing with relevant parameters selection in solar energy prediction problems, in a larger framework of solar energy prediction with neural networks.

Chapter 3

Materials and Methods

3.1 Grouping Genetic Algorithms

The grouping genetic algorithm (GGA) is a class of evolutionary algorithm especially modified to tackle grouping problems, i.e., problems in which a number of items must be assigned to a set of predefined groups. It was first proposed by Falkenauer [31, 32], who realized that traditional genetic algorithms had difficulties when they were applied to grouping problems. Thus, in the GGA, the encoding, crossover and mutation operators of traditional GAs are modified to obtain a compact algorithm, with a high performance in grouping-based problems. The GGA has been successfully applied to different optimization problems, within different applications [52]-[61], but it has not been applied in Feature Selection for ML algorithms (to our knowledge), up until now.

3.1.1 Encoding in GGAs

The GGA considered in this work follows the classical grouping encoding initially proposed by Falkenauer in [31], i.e. it is a variable-length genetic algorithm. The encoding is carried out by separating each individual in the algorithm into two parts: $\mathbf{c} = [\mathbf{l}|\mathbf{g}]$, the first part is the *element* section, whereas the second part is called the *group* section of the individual. As an example, following our notation, in a solution for a problem with N elements (input variables for a solar prediction problem) and k groups, the individual will have the following aspect:

$$l_1, l_2, \ldots, l_N \mid g_1, g_2, \ldots, g_k$$

Note that l_j represents the group to which j-th predictive variable is assigned, whereas group section keeps a list of tags associated to each of the groups of the solution. In a formal way:

$$l_j = g_i \Leftrightarrow \mathbf{x}_j \in C_i. \tag{3.1}$$

Note also that the length of the element section is fixed for a given problem (equals N), but the group section's length is not fixed, but it varies from one individual to another. Thus the GGA does not need as input parameter the number of groups, but it searches for the best k in terms of the objective function.

As an example to fully clarify the GGA encoding, let us suppose the following individual:

1 3 2 1 4 1 1 2 3 2 1 3 4 2 1 | 1 2 3 4

This individual represents a solution with 4 groups of variables, and the following partition of the input variables: $\{x_1, x_4, x_6, x_7, x_{11}, x_{15}\}$, $\{x_3, x_8, x_{10}, x_{14}\}$, $\{x_2, x_9, x_{12}\}$ and $\{x_5, x_{13}\}$.

3.1.2 Selection operator

Different selection operators can be used in the GGA, since the GGA is not defined in exploitation, but in exploration. we describe here a rank-based wheel selection mechanism, previously used and described in [52], but note that any other selection mechanism such as tournament-based selection would be also appropriate. In rank-based wheel selection, the individuals are first sorted in a list based on their quality. The position of the individuals in the list is called *rank of the individual*, and denoted R_i, $i = 1, \dots, \xi$, with ξ number of individuals in the population of the GGA. We consider a rank in which the best individual x is assigned $R_x = \xi$, the second best y, $R_y = \xi - 1$, and so on. A *fitness* value associated to each individual is then defined, as follows:

$$f_i = \frac{2 \cdot R_i}{\xi \cdot (\xi + 1)} \tag{3.2}$$

Note that these values are normalized between 0 and 1, depending on the position of the individual in the ranking list. It is important to note that this rank-based selection mechanism is *static*, in the sense that probabilities of survival (given by f_i) do not depend on the generation, but on the position of the individual in the list. As a small example, consider a population formed by 5 individuals, in which individual 1 is the best quality one ($R_1 = 5$), individual 2 the second best ($R_2 = 4$), and so on. In this case, the fitness associated to the individuals are $\{0.33, 0.26, 0.2, 0.13, 0.06\}$, and the associated intervals for the roulette wheel are $\{0 - 0.33, 0.34 - 0.6, 0.61 - 0.8, 0.81 - 0.93, 0.94 - 1\}$.

This process of selection is usually performed with replacement, i.e., a given individual could be selected several times as one of the parents, however, individuals in the crossover operator must be different.

3.1.3 Crossover operator

The crossover operator defined as the current standard in the GGA is a version of the one initially proposed by Falkenauer in [31]. The process follows a two parents – one offspring scheme, with the following steps:

a. First, two individuals are randomly selected, and two crossing points are chosen in their group part.

b. Insert the elements belonging to the selected groups of the first individual into the offspring.

c. Insert the elements belonging to the selected groups of the second individual into the offspring, if they have not been assigned by the first individual.

d. Randomly complete the elements not yet assigned with elements from the current groups.

e. Remove empty clusters, if any.

f. Modify the labels of the current groups in the offspring in order to numerate them from 1 to k.

Figure 3.1 shows an example of the standard crossover procedure for the GGA. The probability of crossover should be high in the first stages of the algorithm, and moderate in the last ones in order to properly explore the search space. Thus, an adaptive crossover probability could be implemented, defined in the following way:

$$P_c(j) = P_{ci} + \frac{j}{TG}(P_{ci} - P_{cf}) \tag{3.3}$$

where $P_c(j)$ is the crossover probability used in a given generation j, TG stands for the total number of generations of the algorithm, and P_{ci} and P_{cf} are the initial and final values of probability considered, respectively.

3.1.4 Mutation operator

Mutation operator includes small modifications in each individual of the population with a low probability, in order to explore new regions of the search space and also escape from local optima when the algorithm is near convergence. In this case, there can be different procedures which lead to good mutation operators. We present here two different possible mutation operators for a GGA in a Feature Selection problem:

- Mutation by group splitting: it consists of splitting a selected group into two different ones. The samples belonging to the original group are assigned to the new clusters with equal probability. Note that one of the new generated groups will keep its label in the group section of the individual, whereas the other will be assigned a new label $(k+1)$. The selection to the initial group to be split is carried out depending on the clusters' size, with more probability of split given to larger clusters.

  As an example, we illustrate an application of this operator in the final individual from the crossover operator (Figure 3.1), in the case when the initial group chosen to be split is cluster 1:

  2 2 1 3 3 4 4 5 2 1 4 2 5 1 4 | 1 2 3 4 5

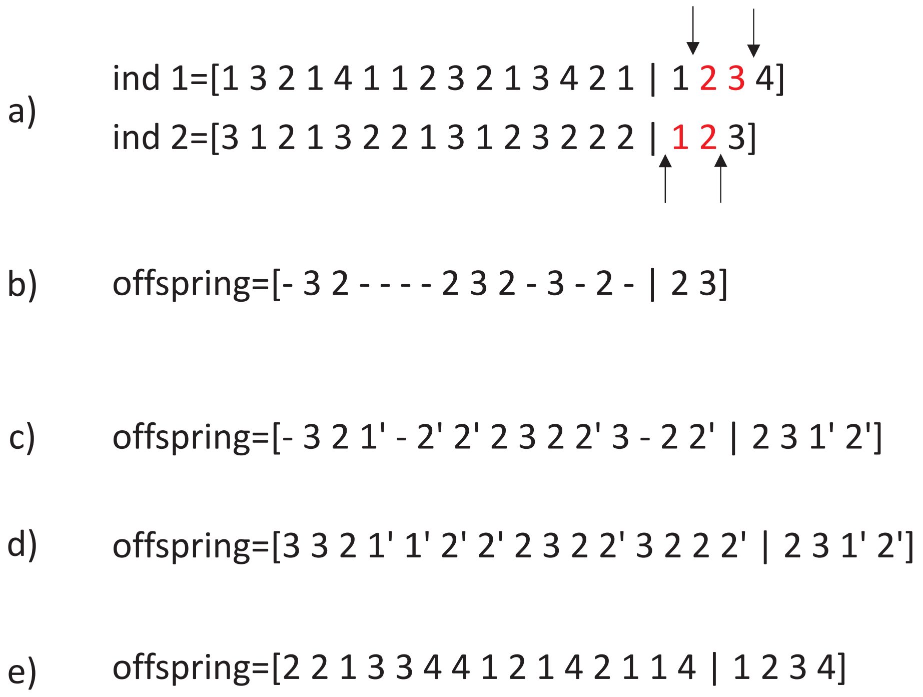

a)
ind 1=[1 3 2 1 4 1 1 2 3 2 1 3 4 2 1 | 1 2 3 4]
ind 2=[3 1 2 1 3 2 2 1 3 1 2 3 2 2 2 | 1 2 3]
b)
offspring=[- 3 2 - - - - 2 3 2 - 3 - 2 - | 2 3]
c)
offspring=[- 3 2 1' - 2' 2' 2 3 2 2' 3 - 2 2' | 2 3 1' 2']
d)
offspring=[3 3 2 1' 1' 2' 2' 2 3 2 2' 3 2 2 2' | 2 3 1' 2']
e)
offspring=[2 2 1 3 3 4 4 1 2 1 4 2 1 1 4 | 1 2 3 4]

Figure 3.1: Outline of the grouping crossover implemented in the GGA.

- Mutation by groups merging: it consists of merging two existing groups, randomly selected, into one. As in mutation by group splitting, the probability of choosing the clusters depends on their size. In order to illustrate this mutation, again an example in the final individual from the crossover operator (Figure 3.1) is given. In this case, let us suppose that the selected groups are clusters 2 and 4:

  2 2 1 3 3 2 2 1 2 1 2 2 1 1 2 | 1 2 3

Similarly to the crossover case, we can also consider an adaptive version of the probability of applying the mutation operators described above. Note that the two mutation operators could be applied in a serial fashion (one after the other), with independent probabilities of application, or just on their own. In this case, probability of mutation should be smaller in the first generations of the algorithm and larger in the last ones, in order to have more opportunities to escape from local minimums in the last stages of the evolution:

$$P_m(j) = P_{mi} + \frac{j}{TG}(P_{mf} - P_{mi}) \tag{3.4}$$

where $P_m(j)$ is the probability of mutation used in a given generation j, TG stands for the total number of generations of the algorithm, and P_{mf} and P_{mi} are the final and initial values of probability considered, respectively.

3.1.5 Replacement and elitism

In the proposed GGA, the population at a given generation $j+1$ is obtained by replacement of the individuals in the population at generation j, through the application of the selection, crossover, and mutation operators described above. An elitist scheme is also applied, so the best individual in generation j is automatically passed on to the population of generation $j+1$, ensuring that the best solution encountered so far in the evolution is always kept by the algorithm.

3.2 The Coral Reefs Optimization Algorithm

The Coral Reefs Optimization algorithm (CRO) is an evolutionary-type algorithm proposed in [99], which is based on simulating the corals' reproduction and coral reefs' formation processes. It has been successfully applied to a number of different applications and optimization problems [100]-[65]. Basically, the CRO is based on the artificial modeling of a coral reef $\mathcal{R}$, consisting of a $n \times m$ grid. We assume that each square *(i,j)* of $\mathcal{R}$ is able to allocate a coral τ_{ij} (candidate solution to the problem, called as $\mathbf{x}$ in the problem's statement above). The CRO algorithm is first initialized at random by assigning some squares in $\mathcal{R}$ to be occupied by corals (i.e. solutions to the problem) and some other squares in the grid to be empty, i.e. holes in the reef where new corals can freely settle and grow in the future. The rate between free/occupied squares in $\mathcal{R}$ at the beginning of the algorithm is denoted as $\rho \in \mathbb{R}(0,1)$ and referred to as initial occupation factor. Each coral is labeled with an associated *health* function $f(\tau_{ij}) : \mathcal{A} \to \mathbb{R}$ that corresponds to the problem's objective function. The CRO is based on the fact that the reef will evolve and develop as long as healthier or stronger corals (which represent better solutions to the problem at hand) survive, while less healthy corals perish.

After the reef initialization described above, the phase of reef formation is artificially simulated. This phase consists of α iterations: at each of such iterations the corals' reproduction in the reef is emulated by applying different operators and processes as described in Algorithm 3: a modeling of corals' sexual reproduction (broadcast spawning and brooding). After the reproduction stage, the set of formed larvae (namely, newly produced solutions to the problem) attempts to find a place on the reef to develop and further reproduce. This deployment may occur in a free space inside the reef (hole), or in an occupied location, by fighting against the coral currently settled in that place. If larvae are not successful in locating a place to settle after a number of attempts, they are considered as preyed by animals in the reef. The coral builds a new reef layer in every iteration.

We detail here the specific definition of the different operators that form the classical CRO algorithm:

1. **Sexual reproduction**: The CRO model implements two different kinds of sexual reproduction: external and internal.

Algorithm 1 Pseudo-code for the CRO algorithm

```
Require: Valid values for the parameters controlling the CRO algorithm
Ensure: A single feasible individual with optimal value of its fitness
Initialize the algorithm
for each iteration of the simulation do
    Update values of influential variables: predation probability, etc.
    Sexual reproduction processes (broadcast spawning and brooding)
    Settlement of new corals
    Predation process
    Evaluate the new population in the coral reef
end for
Return the best individual (final solution) from the reef
```

(a) **External sexual reproduction** or *broadcast spawning*: the corals eject their gametes to the water, from which male-female couples meet and combine together to produce a new larva by sexual crossover. In Nature, some species are able to combine their gametes to generate mixed polyps even though they are different from each other. In the CRO algorithm, external sexual reproduction is applied to a usually high fraction F_b of the corals. The couple selection can be done uniformly at random or by resorting to any fitness proportionate selection approach (e.g. roulette wheel). In the original version of the CRO, standard crossover (one point or two-points) are applied in the broadcast spawning process.

(b) **Internal sexual reproduction** or *brooding*: CRO applies this method to a fraction $(1 - F_b)$ of the corals in the reef. The brooding process consists of the formation of a coral larva by means of a random mutation of the brooding-reproductive coral (self-fertilization considering hermaphrodite corals). The produced larvae is then released out to the water in a similar fashion than that of the larvae generated through broadcast spawning.

2. **Larvae settlement**: once all larvae are formed at iteration k through reproduction, they try to settle down and grow in the reef. Each larva will randomly attempt at setting in a square (i, j) of the reef. If the location is empty (free space in the reef), the coral grows therein no matter the value of its health function. By contrast, if another coral is already occupying the square at hand, the new larva will set only if its health function is better than the fitness of the existing coral. We define a number of attempts $\mathcal{N}_{att}$ for a larva to set in the reef: after $\mathcal{N}_{att}$ unsuccessful tries, it will not survive to following iteration.

3. **Depredation**: corals may die during the reef formation phase of the reef. At the end of each iteration, a small number of corals can be preyed, thus liberating space in the reef for the next iteration. The depredation operator is applied under a very small probability

P_d, and exclusively to a fraction F_d of the worse health corals.

Figure 3.2 illustrates the flowchart diagram of the CRO algorithm, with the different CRO phases (reef initialization and reef formation), along with all the operators described above.

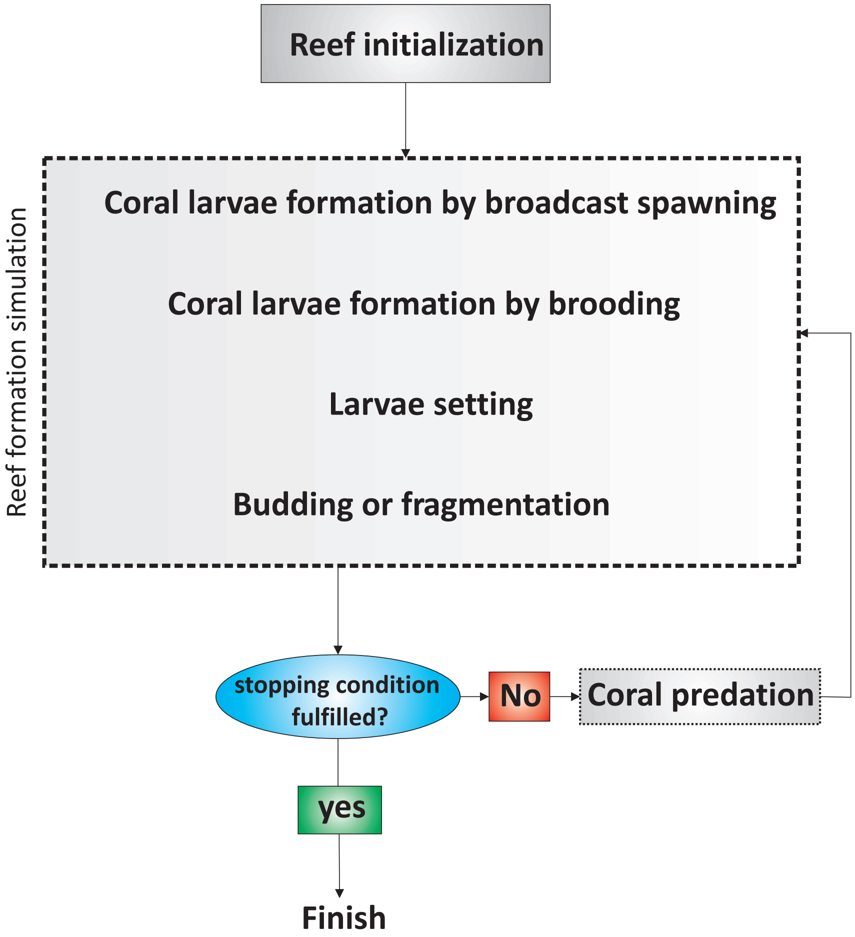

Reef initialization
Reef formation simulation
Coral larvae formation by broadcast spawning
Coral larvae formation by brooding
Larvae setting
Budding or fragmentation
stopping condition fulfilled?
No
Coral predation
yes
Finish

Figure 3.2: Flowchart diagram of the original CRO algorithm.

3.3 Advanced CRO models

The basic CRO can be improved to obtain stronger versions of the meta-heuristic, based on alternative processes that occur in coral reefs. We describe here three different modifications of the CRO algorithm, which improves the performance of this approach in specific applications. First, we describe the CRO with species (CRO-SP), which helps tackle optimization problems with variable length encodings. It is also useful for managing different encodings of problems within the same population, obtaining a competitive co-evolution algorithm. The second CRO version we present here is the CRO with substrates layer (CRO-SL). It has been useful to obtain a competitive co-evolution algorithm in which different searching procedures are applied to one optimization problem within a single population. These model can be either exploration models, repairing mechanisms, etc., and the only pre-requisite is that the objective function to evaluate corals in the reef must be the same for the different models considered. Finally, we show how the CRO can be easily modified to obtain a multi-objective version of the algorithm.

3.3.1 CRO with species (CRO-SP)

The first modification of the CRO consists in considering different coral species within a single coral community. The objective of this modification is that each coral species represents a different model (or its hyper-parameters) out of T possible models. In this context, *model* is generic, so it can represent either a different encoding for the problem, a different way of calculate the objective function, etc. Specifically, the CRO with species is a new powerful way of managing optimization problems with variable encodings. In this case, each species will represent a different encoding length, and the idea is that only corals of the same species can reproduce in the broadcast spawning operator. Note however, that all the models compete together in the larvae setting, since the objective function in all cases should be the same for all the species.

The CRO with species was first introduced in [104] as a methodology to deal with a Model Selection Problem, in an application of total energy consumption prediction in Spain. In [104] each species represents a different way of calculating the total energy consumption (a different model), and the idea was to obtain a competitive co-evolution approach that obtained th1e best possible model in addition to alternative parameters such as the best prediction variables to feed the prediction model. Note that the CRO with species can be used to evolve a competition of different regressions for a given problem, for example neural networks, support vector machines, etc., in which the CRO encodes the parameters of each regressor. Since the concept of *species* is open, it can be used to compare different encodings for a given problem (binary, integer, real, structures, etc.), in which each species corresponds to a given encoding.

Algorithm 2 shows an outline of the CRO containing multiple species. Note that the competition among species will produce emerging behavior, so the best model (species) eventually will dominate, and will occupy the majority of spaces in the reef.

3.3.2 CRO with Susbstrate Layer (CRO-SL)

The original CRO algorithm is based on the main processes of coral reproduction and reef formation that occur in nature. However, there are many more interactions in real reef ecosystem that can be also modelled and incorporated to the CRO approach to improve it. For example, different studies have shown that successful recruitment in coral reefs (i.e., successful settlement and subsequent survival of larvae) depends on the type of substrate on which they fall after the reproduction process [16]. This specific characteristic of coral reefs was first included in the CRO

Algorithm 2 Pseudo-code for CRO algorithm with species

```
Require: Valid values for CRO parameters.
Ensure: The best model out of $T$ possible.
Algorithm initialization ($T$ different species)
for each iteration of the CRO do
    Update values of influential variables: mortality probability and the probability of asexual reproduction
    Asexual reproduction (budding or fragmentation)
    Sexual reproduction 1 (broadcast spawning, only same species can reproduce)
    Sexual reproduction 2 (brooding, only same species can reproduce)
    Settlement of new larvae (competition among species)
    Mortality process
    Evaluate the new population in the reef (with the specific model given for each species)
end for
```

in [104], in order to solve different instances of the Model Type Selection Problem for energy applications. The CRO with substrates is a general approach: it can be defined as an algorithm for competitive co-evolution, where each substrate layer represents different processes (different models, operators, parameters, constraints, repairing functions, etc.). This idea of CRO with substrate layers, was extended as a fully competitive co-evolution search mechanism in [101], where each substrate layer represents a different exploration mechanism. In [105] the interested reader can find more details on alternative co-evolution versions of the CRO algorithm. In this section we describe the main ideas of the CRO-SL as co-evolution search algorithm.

The inclusion of substrate layers in the CRO can be done, in a general way, in a straightforward manner: we redefine the artificial reef considered in the CRO in such a way that each cell of the square grid $\mathcal{R}$ representing the reef is now defined by 3 indexes (i, j, t), where i and j stand for the cell location in the grid, and index $t \in T$ defines the substrate layer, by indicating which structure (model, operator, parameter, etc.) is associated with the cell (i, j). Each coral in the reef is then processed in a different way depending on the specific substrate layer in which it falls after the reproduction process. Note that this modification of the basic algorithm does not imply any change in the corals' encoding. When the CRO-SL is focused on improving the searching capabilities of the classical CRO approach, each substrate layer is defined as a different implementation of an exploration procedure. Thus, each coral will be processed in a different way in the reproduction step of the algorithm. Figure 3.3 shows an example of the CRO-SL, with five different substrate layers. Each one is assigned to a different exploration process, Harmony Search based, Differential Evolution, Gaussian Mutation, M-Points Crossover or 2-Points Crossover. Of course this is only an example and any other distribution of search procedures can be defined in the algorithm. In the specific CRO-SL tested in this paper, each substrate layer only affects to the calculation of the larvae coming from the broadcast spawning process, whereas we have considered the same brooding procedure for all the corals in the reef.

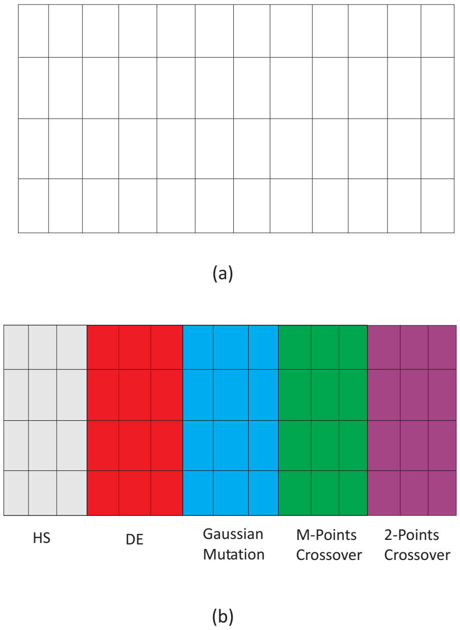


Figure 3.3: Example of CRO-SL and comparison with the original reef in the CRO; (a) Reef considered in the original CRO; (b) Reef in the CRO-SL, where five substrate layers associated with the broadcast spawning process have been considered (Harmony Search (HS), Differential Evolution (DE), Gaussian Mutation, M-Points Crossover and 2-Points Crossover).

There are some important remarks that can be done regarding the CRO-SL approach. First, note that the original CRO is a meta-heuristic based on exploitation of solutions, and leaves the specific exploration open (in the same manner as, for example, Simulated Annealing [59]). This way, the CRO-SL can be seen as a generalization of the original CRO, that does not modify the dynamics of the algorithm (so it can be still outlined following Algorithm 3). The only difference is the specific implementation of the broadcast spawning procedure, which now depends on the specific substrate to which the coral is associated. Second, as has been previously mentioned, the CRO-SL can be seen as a competitive co-evolution procedure. The CRO-SL is a general procedure to co-evolve different models, operators, parameter values, etc., with the only requisite that there is only one health function defined in the algorithm. In this sense, note that the CRO-SL makes a competitive co-evolution of different searching models or patterns within one population of solutions. Finally, note that this approach has been successfully applied to a large number of previous applications, such as battery scheduling and topology design in micro-grids [106, 54], image processing and registration [1, 9], WiFi channel assignment [19], data clustering [117], antennas design [108], Feature Selection problems [125], climate data field reconstruction [107], layout problems [36] or vibration cancellation problems in buildings [95, 18], among others.

3.3.3 Multi-objective CRO

The last modification of the CRO revised here is an adaptation to multi-objective problems, firstly introduced in [102, 103]. In fact, it is a very easy task starting from the basic CRO approach, and only the larvae setting process of the algorithm must be modified: once all the larvae from broadcast spawning and brooding have been produced, they start the setting process one by one, trying to established themselves into the reef. When an existing coral occupies a given position in the reef that is tried by a larva, a fight for the space occurs. In the multi-objective version of the CRO (MO-CRO), this fight for the space is based on domination of solutions. Let us call Ξ_A to the coral currently occupying a given location on the reef, and Ξ_B the larva challenging for the space. In the MO-CRO, Ξ_B wins the fight (and occupies the place of Ξ_A) if and only if $\Xi_A \prec \Xi_B$, where $\prec$ stands for the dominance operation see [103]. In any other case, Ξ_A wins, and the challenging larva either tries another place in the reef, or die, depending on its current κ value. Note that in case of equivalency between solutions ($\Xi_A \equiv \Xi_B$), the current solution in the reef is maintained. A second adaptation is needed in order to provide diversity to the reef: In the MO-CRO, a fix number μ of corals with the same value in all objective functions is allowed in the population. After the larvae setting process, the number of corals in the reef with the same value in all objectives is obtained, let us call it β, and if $\mu < \beta$, $(\beta - \mu)$ randomly chosen corals are depredated. This adaptation will be known as Extreme Depredation Operator (EDO). The pseudo-code of Algorithm 3 describes the MO-CRO process.

3.4 Multi-Layer Perceptrons

A multi-layer perceptron is a particular class of artificial neural network, which has been successfully applied to solve a large variety of nonlinear problems, mainly classification and regression tasks [42, 14]. The multi-layer perceptron consists of an input layer, a number of hidden layers, and an output layer, all of which are basically composed by a number of special processing units called *neurons*. All the neurons in the network are connected to other neurons by means of weighted links (Figure 3.4). In the multi-layer perceptron, the neurons within a given layer are connected to those of other layers. The values of these weights are related to the ability of the multi-layer perceptron to learn the problem, and generalize from a sufficiently long number of examples. The process of assigning values to these weights from labelled examples is known as the training process of the perceptron. The adequate values of the weights minimizes the error between the output given by the multi-layer perceptron and the corresponding expected output in the training set. The number of neurons in the hidden layer is also a parameter to be optimized [42, 14].

The input data for the multi-layer perceptron consists of a number of samples arranged as input vectors, $\mathbf{x}=\{x_1, \ldots, x_N\}$. Once a multi-layer perceptron has been properly trained,

Algorithm 3 Pseudo-code for the MO-CRO algorithm

```
Require: Valid values for the parameters controlling the CRO algorithm
Ensure: A feasible optimal pareto front of solutions to the optimization problem
Initialize the algorithm: Set values for $P_d$, $\mu$, $F_a$, $F_b$, $F_d$.
for each iteration $k$ of the simulation do
    ∘ Sexual crossover process (Brooding):
    for each brooding coral $\Xi_{i,j}$ do
        $\Xi_{i,j} \to \Xi^m$
    end for
    ∘ Sexual crossover process (Broadcast Spawning):
    for couples of broadcast spawning corals $\Xi_{i,j}$ and $\Xi_{k,j}$ do
        $\Xi_{i,j} + \Xi_{k,j} \to \Xi^b$
    end for
    ∘ Asexual crossover process (Fragmentation):
    $F_a$ of the best corals in the reef duplicate and are mutated $\Xi_{i,j} \to \Xi^c$
    ∘ Settlement of new corals (dominant solutions prevail in the reef): $\{\Xi^m, \Xi^b, \Xi^b, \Xi^c\} \to \Xi_{i,j}$
    ∘ Predation process (EDO operator):
    for all corals in the reef do
        eliminate copies of corals until just $\mu$ remain in the reef.
    end for
    ∘ Calculate corals' health $f(\Xi_{i,j}) : \mathcal{I} \to \mathbb{R}$,
end for
Return the pareto front formed in the coral reef
```

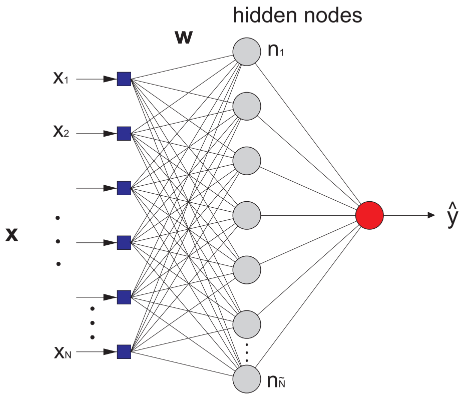


Figure 3.4: Structure of a multi-layer perceptron neural network, with one hidden layer.

validated and tested using an input vector different from those contained in the database, it is able to generate a proper output y. The relationship between the output and the input signals of a neuron is

$$\vartheta = \varphi\left(\sum_{j=1}^{n} w_j x_j - \theta\right), \tag{3.5}$$

where ϑ is the output signal, x_j for $j = 1, \ldots, n$ are the input signals, w_j is the weight associated with the j-th input, and θ is a threshold [42, 14]. The transfer function φ is usually considered as the logistic function,

$$\varphi(x) = \frac{1}{1 + e^{-x}}. \tag{3.6}$$

The well-known Stochastic Gradient Descent (SGD) algorithm is often applied to train the multi-layer perceptron [17]. One the the main issues with this and other adaptive algorithms for training multi-layer perceptrons is their high computational cost, which makes their application in hybrid approaches difficult. In the next section we present the Extreme Learning Machine, a fast training procedure for multi-layer perceptrons which solves this point.

3.5 Extreme-Learning Machines

An Extrem Learning Machi (ELM) [44] is fast training method for neural networks, which can be applied to feed-forward perceptron structures (Figure 3.4). In the ELM, the network weights of the first layer are set at random, and after this, a pseudo-inverse of the hidden-layer output matrix is obtained. This pseudo-inverse is then used to obtain the weights of the output layer which fits best with the objective values. The advantage of this method is not only that it it is extremely fast, but also that it obtain competitive results versus other established approaches, such as classical training for multi-layer perceptrons, or support-vector-regression algorithms, etc. The universal-approximation capability of the ELM have been proven in [138].

The ELM algorithm can be summarized by considering a training set,

$$\mathbb{T} = (\mathbf{x}_i, \boldsymbol{\vartheta}_i) | \mathbf{x}_i \in \mathbb{R}^n, \boldsymbol{\vartheta}_i \in \mathbb{R}, i = 1, \cdots, l,$$

an activation function $g(x)$, and a given number of hidden nodes $(\tilde{N})$, and applying the following steps:

1. Randomly assign input weights $\mathbf{w}_i$ and the bias b_i, where $i = 1, \cdots, \tilde{N}$, using a uniform probability distribution in $[-1, 1]$.

2. Calculate the hidden-layer output matrix $\mathbf{H}$, defined as follows:

$$\mathbf{H} = \begin{bmatrix} g(\mathbf{w}_1\mathbf{x}_1 + b_1) & \cdots & g(\mathbf{w}_{\tilde{N}}\mathbf{x}_1 + b_{\tilde{N}}) \\ \vdots & \cdots & \vdots \\ g(\mathbf{w}_1\mathbf{x}_l + b_1) & \cdots & g(\mathbf{w}_{\tilde{N}}\mathbf{x}_N + b_{\tilde{N}}) \end{bmatrix}_{l\times\tilde{N}}. \tag{3.7}$$

3. Finally, calculate the output weight vector β as follows:

$$\beta = \mathbf{H}^{\dagger}\mathbf{T}, \tag{3.8}$$

where $\mathbf{H}^{\dagger}$ is the Moore-Penrose inverse of the matrix $\mathbf{H}$ [44], and $\mathbf{T}$ is the training output vector, $\mathbf{T} = [\boldsymbol{\vartheta}_1, \cdots, \boldsymbol{\vartheta}_l]^T$.

Note that the number of hidden nodes ($\tilde{N}$) is a free parameter to be set before the training of the ELM algorithm, and must be estimated for obtaining good results by scanning a range of $\tilde{N}$.

Part III

Part II: Experiments and results

Chapter 4

A GGA-ELM approach for global solar radiation prediction from numerical weather model inputs

4.1 Introduction

The objective of this chapter is twofold: first, we consider a problem of global solar radiation prediction from numerical weather models, specifically the WRF meso-scale model [111], which illustrate the case of input data from NWM. Thus, the WRF provides a prediction of atmospheric variables at different pressure levels in a given zone, that will be used as inputs in a prediction system to estimate the global solar radiation at a different point. The second contribution of the chapter is the development of a hybrid grouping genetic algorithm – ELM (GGA-ELM) algorithm to carry out this global solar radiation prediction. As previously explained, the GGA proposed will perform a process of feature selection, focused on filtering the best features from the WRF model to do the prediction, whereas the ELM approach will do the final prediction of the global solar radiation at a given point, using the features selected by the GGA.

In this chapter, we discuss in detail the proposed algorithm, giving some variants of its dynamics, that lead to different performances. With this algorithmic framework in mind, we then tackle a number of subproblems related to global radiation prediction: the first problem considered consists of predicting the solar radiation registered in a given point $\mathcal{P}$ at time $t + x$ (for $x = 0, \ldots, 3$), using as predictive variables the set $\mathcal{V}$, or any subset of it. Note that when $x = 0$, the problem is known as *statistically downscaling* the solar radiation prediction of model $\mathcal{M}$ to point $\mathcal{P}$. The ultimate goal of this approach for $x = 0$ (Subproblem 1) is to evaluate what features (predictive variables) from the NWM are useful for this prediction. Note that for $x > 0$ (Subproblem 2) we are evaluating the prediction performance of the system only using as predictive variables the outputs of the WRF. Finally, we tackle in this paper a forecasting problem that takes into account data from the NWM and also objective variable data measured at the measuring station considered (Subproblem 3). This last subproblem uses the best predictive variables set found in Subproblem 1. Results for all these subproblems using real data from Toledo's radiometric station (Spain) will be discussed in the experimental part of the chapter.

4.2 Problem formulation

The solar radiation problem considered can be stated in the following way: Let $\mathcal{P}$ be a given location of the Earth's surface where the global solar radiation ($\mathcal{I}_t$) must be predicted ($\hat{\mathcal{I}}_t$), at a given time t. To do this, let us consider the output, $\mathcal{V}$, of a numerical weather meso-scale model $\mathcal{M}$, in a number M of nodes, consisting of the prediction at time t for N atmospheric variables, $\mathcal{V} = (\varphi_{11}, \dots, \varphi_{1N}, \varphi_{21}, \dots, \varphi_{2N}, \dots, \varphi_{M1}, \dots, \varphi_{MN})$, as shown in Figure 4.1.

Note that $\mathcal{M}$ may provide an atmospheric variable at the ground level, or at the ground level and also at different pressure levels. In the latter, each pressure level is considered as a different variable n at node m, φ_{mn}.

The problem under study were analysed over 3 different points of view, leading to 3 kind of subproblems:

1. Subproblem 1 deals with the prediction of the global solar radiation registered in $\mathcal{P}$ at time t, using as predictive variables the set $\mathcal{V}$, or any subset of it. This type of problems is usually known in other works as *statistically downscaling* the solar radiation prediction of model $\mathcal{M}$ to point $\mathcal{P}$.The ultimate goal of this approach is to evaluate what features (predictive variables) from the Numerical Model are useful for this prediction.

2. Subproblem 2 increases the forecast horizon, predicting the global solar radiation in $\mathcal{P}$ at time $t+x$ (for $x = 1, 2, \dots, \mathcal{X}$), considering the set $\mathcal{V}$ (or any subset) as predictive variables.

3. Subproblem 3 analyzes a forecasting problem at $\mathcal{P}$ considering previous radiation values. For this purpose, the best set of features found by the GGA-ELM in Subproblem 1, i.e. $\mathcal{V}^* = (\varphi_1^*, \dots, \varphi_{\mathcal{K}}^*)$, where $\mathcal{K}$ is the number of optimal features obtained by the GGA-ELM approach, have been used. In addition, we consider objective variable data measured at point $\mathcal{P}$ for previous time tags ($\mathcal{I}_{t-z}$), for $z = 1, \dots, \mathcal{Z}$), in the process of $\hat{\mathcal{I}}_{t+x}$ forecasting, for $x = 1, 2, \dots, \mathcal{X}$.

4.2.1 Location of the measurement station and objective variable data

This work predicts the global solar radiation at a given location $\mathcal{P}$ that pinpoints a Meteorological State Agency of Spain (AEMET) station sited in Toledo (39° 53' 5"N, 4° 02' 43"W). Toledo's measuring station is located in the South Plateau of the Iberian Peninsula (See Figure 4.2), around 75 km south of Madrid (the capital of Spain) at an altitude of 515 m.

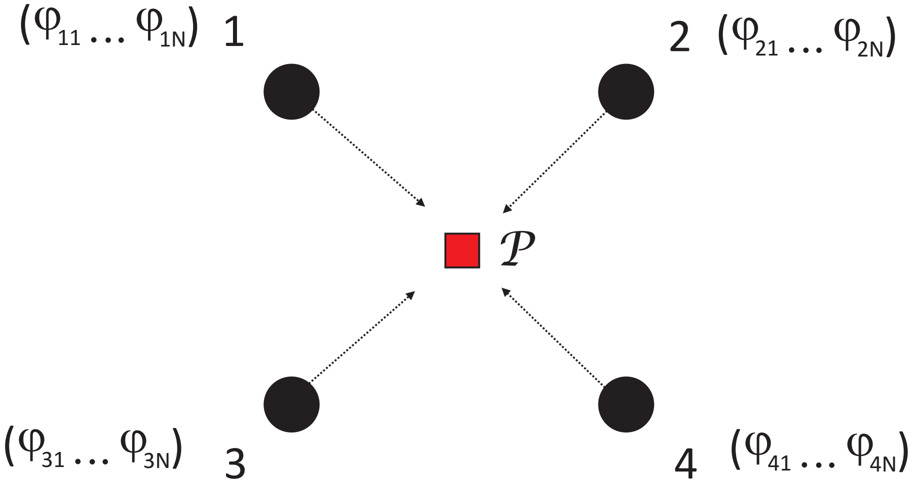


Figure 4.1: Solar radiation prediction scheme used in this work for $M = 4$.

According to AEMET's Climate Summary Guide (1981 - 2010), Toledo has an annual mean temperature of 15.8°C, dry summers, an annual mean precipitation of 342.2 mm, and an annual mean water vapour tension of 10.8 hPa. In the light of these figures and according to the Köppen climate classification, Toledo could be classified as a Csa climate (Interior Mediterranean: Mild with dry, hot summer). Regarding cloud cover, 27.7% annual days are categorized as cloud free, 53.2% present sky conditions categorized as few or scattered clouds, and the remaining 19.1% are categorized as broken or overcast.

As objective variable data to train and test the algorithms, we consider one year of hourly global solar radiation data (from May 1st, 2013 to April 30th, 2014) collected at Toledo's measuring station. To measure global solar radiation, a Kipp & Zonen CMP11 Pyranometer is used. All radiation measurements gathered by the AEMET are made following the World Meteorological Organization (WMO) standards included in the WMO Guide to Meteorological Instruments and Methods of Observation (2008 edition, updated in 2010).

4.2.2 Model $\mathcal{M}$: the Weather Research and Forecasting model (WRF)

In this experiment we use the well-known Weather Research and Forecasting (WRF) meso-scale model as $\mathcal{M}$ [111]. WRF is an extremely powerful meso-scale numerical weather prediction system designed for atmospheric research and also for operational forecasting needs. The WRF was developed in collaboration by the National Center for Atmospheric Research (NCAR), the National Centers for Environmental Prediction (NCEP), the Forecast Systems Laboratory (FSL), the Air Force Weather Agency (AFWA), the Naval Research Laboratory, the University

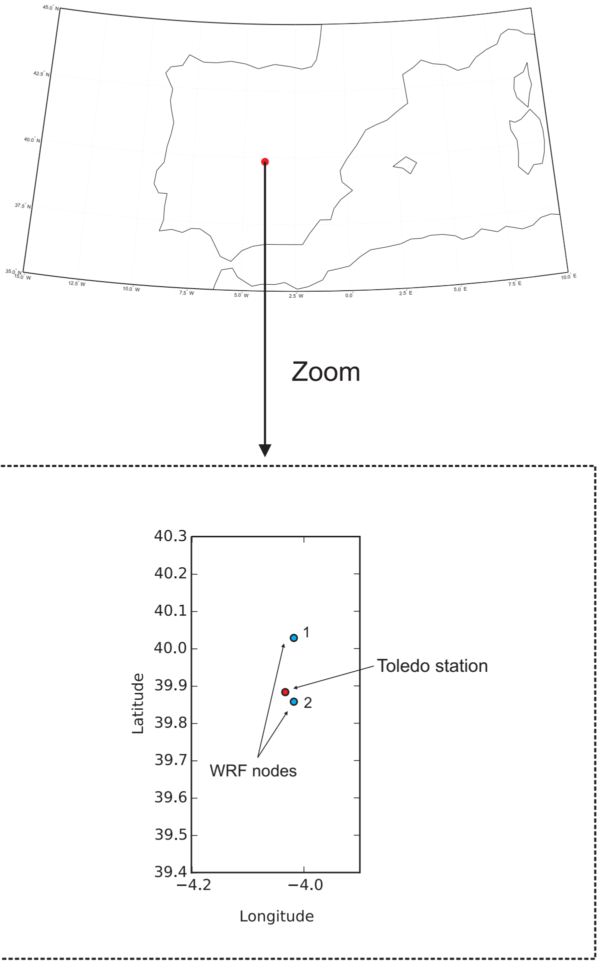


Figure 4.2: Location of the Toledo's measuring station and the $M = 2$ WRF nodes considered for the prediction.

of Oklahoma, and the Federal Aviation Administration (FAA) of the USA. The WRF has been used in a wide range of meteorological [40] and renewable energy applications [20].

In this study, WRF model version 3.6 has been used. It has been run every 12 hours since it was started in 2011. Meteorological data are calculated over a window ranging in latitude from 34° 33' 43"N to 44° 28' 12"N , and in longitude from 4° 25' 12"W to 4° 23' 2"E. In this window, the grid has 99 elements from West to East, and 59 elements from North to South, roughly, each grid element covers 15×30 km^2. Atmospheric values are calculated, in the vertical dimension, at 37 levels above the ground, at ground level, and at four additional levels beneath the surface. The grid type is Arakawa, that is to say that data are calculated at the center of each element, with a 72 seconds time step.

WRF is initialized by data coming from NCEP FNL Operational Global Analysis and works in non-hydrostatic way. The short wave scheme used is that from MM5 shortwave (Dudhia), and the long wave model is the RRTM (Rapid Radiative Transfer Model). A radiation time step of 30 minutes was applied to each radiation domain. The land surface fluxes were obtained by Monin-Obukhov similarity theory, the surface physics was solved by the Unified Noah land surface model and the Planetary Boundary Level (PBL) by means of the Yonsei University (YSU) PBL scheme. The PBL was calculated at every basic time step, and five layers were considered in land surface model. Cumulus retrieval parameters was done by using the new Kain-Fritsch scheme, as in MM5 and Eta/NMM ensemble version, with a time step of 5 minutes.

Finally, micro-physics was carried out by the WSM 3-class scheme and the turbulent diffusion option was to select 2^{nd} order diffusion on model levels. This complements vertical diffusion done by the PBL scheme.

WRF output at two points located at (39° 51'N, 4° 01'W) and (40° 01'N, 4° 01'W), are selected as predictive variables for solar radiation (see Figure 4.2), both points are the 2 nearest WRF's points to the target location. Specifically, Table 4.1 shows the 46 variables considered for each of these points, summing up a total of 92 predictive variables for this problem. They are the following (at different pressure levels):

- OLR: The top of atmosphere outgoing long-wave radiation (W/m^2).

- GLW: The downward long-wave flux at ground surface (W/m^2).

- SWDOWN: The downward short-wave flux at ground surface (W/m^2).

- u: The horizontal wind component in the x direction at different pressure levels (m/s).

- v: The horizontal wind component in the y direction different pressure levels (m/s).

- CLDFRA: The fraction of clouds in each cell. Cloud fraction ranges from 0 (no clouds) to 1 (clouds in a spatial grid cell).

- QVAPOR: The water vapor mixing ratio (in kg/kg). This variable is defined as the ratio of the mass of a water vapor to the mass of dry air.

- T: The temperature (in K) at different levels. Note that this variable is not directly provided by the WRF model. Thus, we have obtained it from the WRF perturbation potential temperature (T') which, in turn, is related with the potential temperature (θ) through the relation $\theta = T' + 300$. Potential temperature is simply defined as the temperature that an unsaturated parcel of dry air would have if brought adiabatically and reversibly from its initial state to a standard pressure, P_0, typically 100000 Pa. Its mathematical expression is as shown in Eq. 4.1, where κ is the Poisson constant.

$$T = \theta\left(\frac{P}{P_0}\right)^{\kappa} \tag{4.1}$$

Table 4.1: Predictive variables used in the experiments (46 variables per node of the WRF model).

variable	units	pressure levels (hPa)
OLR	W/m^2	-
GLW	W/m^2	ground
SWDOWN	W/m^2	ground
u	m/s	ground, 850, 700, 500, 400, 300, 200, 100, 50
v	m/s	ground, 850, 700, 500, 400, 300, 200, 100, 50
CLDFRA	1/0	ground, 850, 700, 500, 400, 300, 200
QVAPOR	kg/kg	ground, 850, 700, 500, 400, 300, 200, 100, 50
Temperature	K	ground, 850, 700, 500, 400, 300, 200, 100, 50

4.3 Specific GGA implemented

The GGA proposed for feature selection in this problem of solar radiation prediction follows the encoding structure defined in Section 3.1.1. Specifically, every individual's assignment part is composed of 92 elements in this case, each corresponding to one of the predictive variables provided by the WRF model, as explained in Section 5.2. Then, the group part is composed of a variable number of elements (groups), as defined in Section 3.1.1.

4.3.1 Genetic operators

In this case, we have considered as selection operator a tournament-based mechanism, similar to the one described in [127], as it has been shown to be one of the most effective selection operators, avoiding super-individuals and performing an excellent exploration of the search space. This operator allows that those individuals which are not the best solutions in our search space to be selected as parents of the next generation in a few percentage of situations. This simple action avoids the appearance of local minimums which introduce noise in our experiment.

Regarding the crossover operator, different versions of the operator described in Section 3.1.3 have been implemented for this problem, and their results obtained compared. For a clearer description, notation used refers to those individuals acting as parents as P_i $(i = 1, 2)$ and those individuals acting as offsprings as O_i $(i = 1, 2)$. The assignment and groups part of a certain individual are referred as A_{P_i} and G_{P_i} (for the parents), or A_{O_i} and G_{O_i} (for the offsprings).

The first crossover operator applied follows the guidelines initially proposed by Falkenauer [31, 32], and described in Section 3.1.3, that leads to a two-parents/one-child mechanism. The process (outlined in Figure 4.3) carried out by this first crossover operator, $\mathcal{C}_1$ hereafter, is the following:

1. Randomly choose two parents from the current population: P_1 and P_2. The offspring individual, O_1, is initialized to be equal to P_2.

2. Randomly select, for the crossover, two points from G_{P_1}. These two cross-points mark down those groups in-between them, and those features assigned to these groups are selected. In the example presented in Figure 4.3, the two crossover points select two groups: group number 1 (G_1) and group number 2 (G_2). Note that, in this case, the features of P_1 belonging to groups G_1 and G_2 are 1, 2, 4, 5, and 6 (marked bold and underlined).

3. Insert those P_1's selected features (in their own positions) in O_1. Then, attach at the end of O_1's group section those new groups inherited from P_1. In the example, it can be seen that the assignment of the features 1, 2, 4, 5 and 6 of O_1 has been inherited from P_1, while the rest of the nodes' assignment has been inherited from P_2.

4. Rename G_{O_1}'s groups to remove duplicates (note that the offspring may have inherited same groups' numbering from both parents). In the example, $G_{O_1} = 1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6$ $1 \quad 2$ is changed to $G_{O_1} = 1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6 \quad 7 \quad 8$. Therefore, A_{O_1} has to be modified accordingly.

5. Remove, empty groups in O_1, if present. In the example considered, it is found that O_1's groups 1, 2, 3, and 6 are empty (there are no features belonging to them), so we can eliminate these groups' identification number and rearrange the rest accordingly. The final offspring is then obtained.

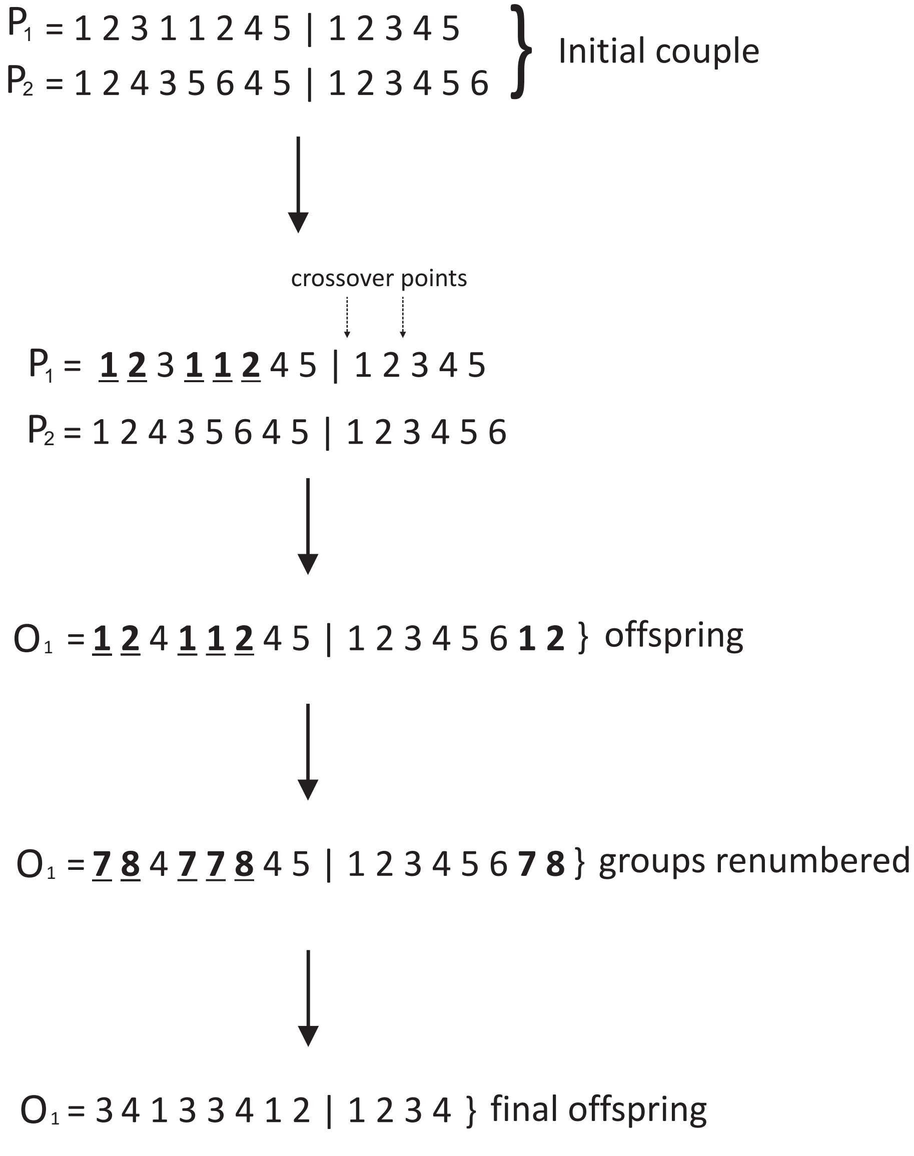

P1 = 1 2 3 1 1 2 4 5 | 1 2 3 4 5
P2 = 1 2 4 3 5 6 4 5 | 1 2 3 4 5 6
Initial couple
crossover points
P1 = 1 2 3 1 1 2 4 5 | 1 2 3 4 5
P2 = 1 2 4 3 5 6 4 5 | 1 2 3 4 5 6
O1 = 1 2 4 1 1 2 4 5 | 1 2 3 4 5 6 1 2
offspring
O1 = 7 8 4 7 7 8 4 5 | 1 2 3 4 5 6 7 8
groups renumbered
O1 = 3 4 1 3 3 4 1 2 | 1 2 3 4
final offspring

Figure 4.3: Outline of the grouping crossover $\mathcal{C}_1$, implemented in the proposed GGA.

In some initial experiments carried out for this problem, we realized that $\mathcal{C}_1$ crossover operator produced a significant increment of the number of groups after a number of generations of the GGA. This situation is due to the type of problem faced: all the groups perform better in the first stages of the algorithm when the number of features decrease on them, and so the algorithm tried to reduce the number of features by artificially increasing the number of groups, adding noise to the fitness calculation process. We tried to correct this issue with this version of the GGA crossover by introducing an alternative grouping crossover mechanism. Accordingly, a two-parents/two-children crossover, $\mathcal{C}_2$, is also introduced. The $\mathcal{C}_2$ process is shown in Figure 4.4, and can be described as follows:

1. Randomly choose two parents from the current population: P_1 and P_2.
2. Randomly select, for the crossover, two points from G_{P_1} and two points from G_{P_2}. For each parent, these cross-points mark down those groups in-between them, and those features assigned to these groups are selected.
3. To build G_{O_1}, use the selected section of G_{P_1}. In the example, P_1's selected groups are G_2 and G_3, resulting in the offsprings group part $G_{O_1} = 2 \quad 3$.
   To build A_{O_1} use the selected features inherited from P_1.
4. If necessary, rename G_{O_1}'s groups so that groups' numbering starts at 1.
5. Randomly allocate among the offspring's groups those blank features. The final first offspring is then obtained.
6. Repeat steps 2 to 5 using the second parent to obtain the second offspring.

We decided to add a new procedure to reduce the number of groups when the experiment grows in number of generations. A variable K which decides the maximum number of groups each individual contains. If after the crossover, the number of groups is bigger than K, the worst groups will be erased and their features randomly added to the rest of the groups.

Regarding mutation operator, a swapping mutation in which two items are interchanged is applied in this problem. Thus, resulting in the assignment of features to different groups. This procedure is carried out with a very low probability (P_m = 0.01), to avoid increasing the random search in the process [32].

4.3.2 Fitness function: ELM output

The fitness function considered for each element (group of features) of the GGA is obtained by using the Root Mean Square Error (RMSE) of the prediction given by the ELM (see Section 4.3.2). RMSE is used in this work instead of other validation metrics, because large forecast errors and outliers are weighted more strongly than smaller errors, as the latter are more tolerable in solar radiation prediction [10, 60].

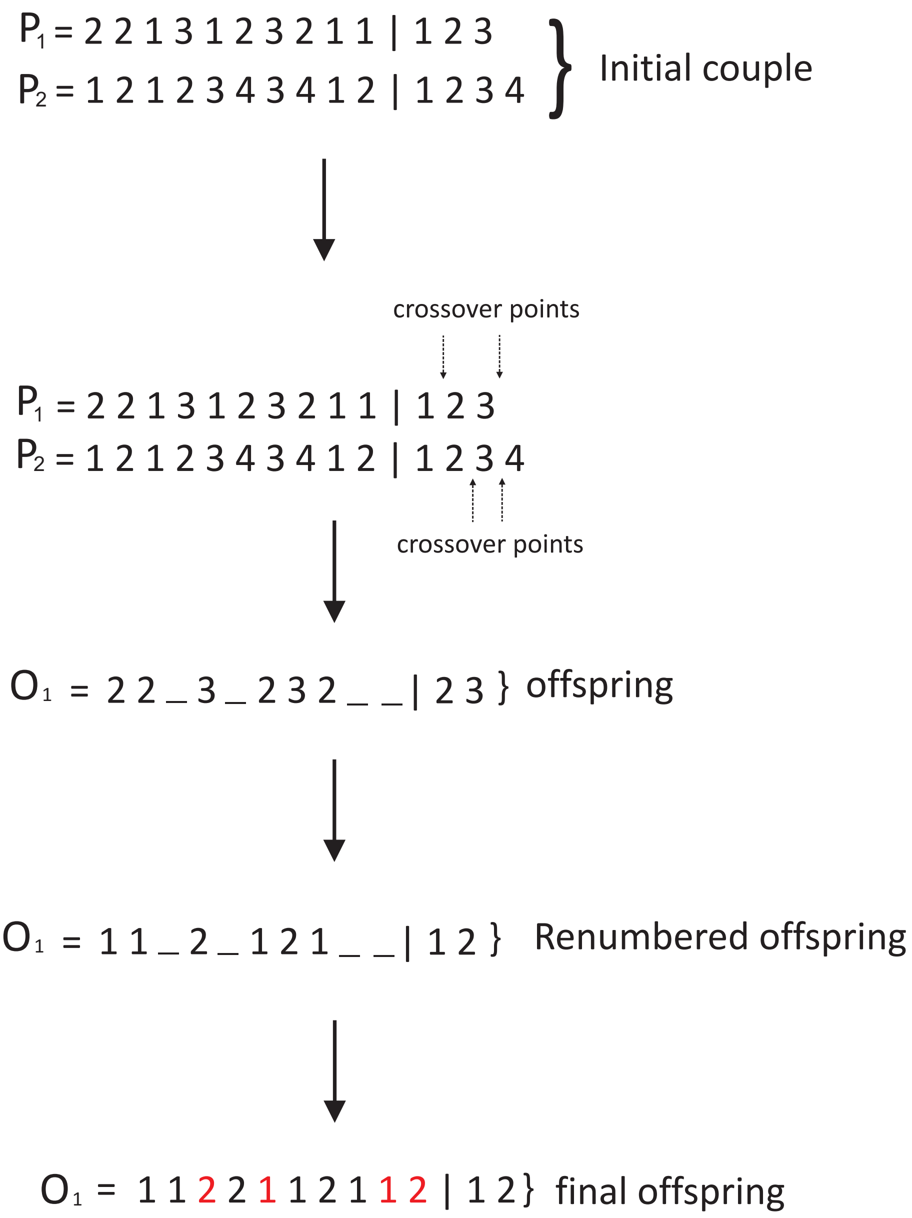

P1 = 2 2 1 3 1 2 3 2 1 1 | 1 2 3
P2 = 1 2 1 2 3 4 3 4 1 2 | 1 2 3 4
Initial couple
crossover points
P1 = 2 2 1 3 1 2 3 2 1 1 | 1 2 3
P2 = 1 2 1 2 3 4 3 4 1 2 | 1 2 3 4
crossover points
O1 = 2 2 _ 3 _ 2 3 2 _ _ | 2 3
offspring
O1 = 1 1 _ 2 _ 1 2 1 _ _ | 1 2
Renumbered offspring
O1 = 1 1 2 2 1 1 2 1 1 2 | 1 2
final offspring

Figure 4.4: Outline of the grouping crossover $\mathcal{C}_2$, implemented in the proposed GGA.

The RMSE formula is shown in Equation (5.1), where $\mathcal{I}_t$ stands for the global solar radiation measured at a time t, $\hat{\mathcal{I}}_t$ stands for the global solar radiation estimated by the ELM, and T stands for the number of samples in the test set.

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\hat{\mathcal{I}}_t - \mathcal{I}_t\right)^2} \tag{4.2}$$

Note that RMSE is the only metric used at the ELM's training phase. Nevertheless, to assess the performance of the network (in the test phase), alternative metrics have been also used (see Subsection 4.4.1).

4.3.3 GGA evolution dynamics

Feature selection is performed in this problem of solar radiation prediction with the GGA. The possible 92 predicting variables of the problem are assigned to the different groups defined in the groups' part of the individual. Since each individual is divided into several groups, there are different approaches to calculate the fitness function (given by Equation (5.1)), as all groups may be used in this calculation, just one, and so on. In this work we have analyzed two different evolution dynamics for the GGA:

1. Dynamics $\mathcal{D}_1$: The fitness function given by Equation (5.1) is calculated for all groups in each individual, and the fitness value is assigned by choosing the minimum value for all the groups.

2. Dynamics $\mathcal{D}_2$: The fitness function is also given by Equation (5.1), but in this case we choose to maximize the value of this equation for group G_1, called "junk group".

Note that $\mathcal{D}_1$ is the most intuitive way of the GGA evolution, where the individuals are selected according to the best fitness value obtained for one of their groups. On the other hand, $\mathcal{D}_2$ is completely different: in this case the idea is to concentrate those features that produce a poor performance of the regressor in a given group (group G_1 in this case). At the end of the evolution, the test value is obtained using the features that are not present in the first group of the best individual in the population. Note that $\mathcal{D}_2$ can be improved by eventually removing the worst features out of the total available ones after a number of generations (η). In order to do this, after η generations of the algorithm, we construct a ranking of those features that appear the most in group G_1 of all the individuals in the population. We then set a threshold (th) for the number of times a given feature appears in group G_1, and we remove those features that appear in the ranking over threshold th. The GGA is then re-initialized without considering those features that were removed in the previous step.

Regarding execution/computation time, note that $\mathcal{D}_1$ is heavier than $\mathcal{D}_2$. The reason for this is that the number of groups analyzed per individual and generation are different in each dynamics: $\mathcal{D}_1$ needs to evaluate each group of each individual, whereas $\mathcal{D}_2$ evaluates only one group per individual. For example, in a case of an average of 10 groups per individual, and 50 individuals, $\mathcal{D}_1$ checks 500 groups per generation against 50 groups evaluated by $\mathcal{D}_2$. In this case, this suggests that $\mathcal{D}_2$ could carry out a deeper and faster search.

4.4 Experiments and results

This section summarizes the experiments run to assess the proposed algorithm, together with measures used to evaluate the accuracy of our approach, and to allow comparison with other global solar radiation prediction tools.

Note these experiments consider global solar radiation prediction only for daytime hours (average hourly data from 5 a.m. to 8 p.m.), as night hours present zero irradiance. In order to create training and test sets to evaluate the performance, each day has been divided into several blocks. For each block of 4 hours, a random hour is assigned to the test set, and the remaining ones, in increasing order, to the training set. Thus, guaranteeing that all blocks and all days are represented both in train and test. Finally, this procedure is carried out 10 times, and we then provide average values of error in test (10-fold cross validation).

4.4.1 Forecasting accuracy measures

Following the guidelines given in [10], in order to assess the forecasting accuracy of for global solar radiation prediction, conventional metrics such as RMSE (Equation (5.1)) and Pearson Correlation Coefficient, r^2, have been used in this experiment.

Moreover, to facilitate comparisons between the forecasts developed in this work and other solar forecasts, another useful quality check is to analyze whether the proposed GGA-ELM model performs better or worse than a given reference model [60]. Thus, forecast skill, s, defined by Equation (4.3) has been also considered, where U stands for the uncertainty of solar availability, i.e. the forecasting error of the proposed GGA-ELM model, and V refers to the variability of solar irradiance. Taking into account that this variability can be attributed to cloud cover (mostly stochastic) and solar position (mostly deterministic), it can be referred as the standard deviation of the step-changes of the ratio of the measured solar irradiance to that of a clear-sky solar irradiance. An easier procedure to obtain the forecast skill, and yet a good estimate [60], is to consider that the ratio U/V can be approximated by the $RMSE_{prediction}/RMSE_{persistence\ model}$. Global solar radiation obtained with the persistence model at a point $\mathcal{P}$ and at a time t, will be referred as $\mathcal{I}_t^{Per}$.

$$s = 1 - \frac{U}{V} \tag{4.3}$$

4.4.2 Subproblem 1

Subproblem 1 evaluates what WRF model outputs (the predictive variables) are more useful in the global solar radiation prediction. To evaluate it, we have carried out several experiments to test the proposed hybrid GGA-ELM algorithm. First, the GGA will perform a feature selection out of the 92 possible atmospheric output variables considered. Note, that 92 features are selected by an expert from the whole set of features coming from WRF. Then, the ELM will perform the global solar prediction using those features selected by the GGA in the test set.

The result for global solar radiation prediction with the ELM when no feature selection is performed (tested on all the available features) is $r^2 = 0.9283$ and $RMSE = 85.14\ W/m^2$ (average of the 10-fold cross validation). Note that this is a baseline reference that should be outperformed by the proposed algorithm. Figures 4.5 and 4.6 present the scatter plot and global solar prediction in time, where it can be seen that the prediction fits rather well to the field (measured) data. For a clearer representation, only the first 100 hours of the test output are shown in all time graphs.

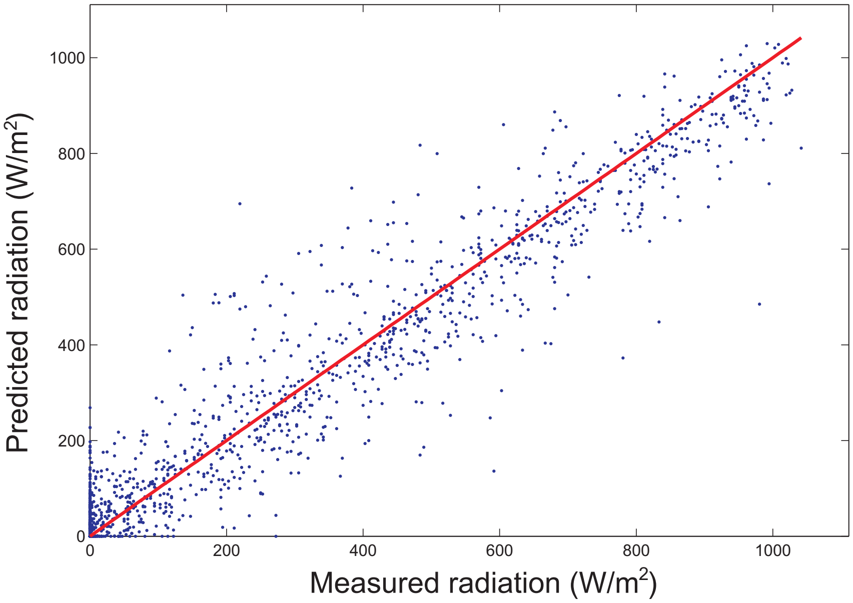


Figure 4.5: Scatter plot of the global solar radiation prediction by the ELM without feature selection.

Now, in order to test the hybrid GGA-ELM approach, several experiments were run showing the results presented in Table 4.2. Note that due to the ELM's output variability, when the hybrid GGA-ELM approach is tested, small increases in the RMSE may occur at certain generations. In order to reduce this variability, the ELM is run γ times ($\gamma = 10$) at each iteration (for both dynamics) and the average RMSE value is used as the individual's fitness value (the minimum RMSE of all possible groups in each individual for dynamics $\mathcal{D}_1$, or the maximum RMSE in the individual's group G_1 for dynamics $\mathcal{D}_2$). Therefore, every generation in $\mathcal{D}_1$, the algorithm runs γ ELMs for each group at each individual ($\gamma \cdot N_{Groups}$), while every generation in $\mathcal{D}_2$ only runs γ ELMs for group G_1 at each individual, obtaining a big reduction in terms of time computing.

The first experiments compare the two crossover operators presented in Section 4.3.1: $\mathcal{C}_1$

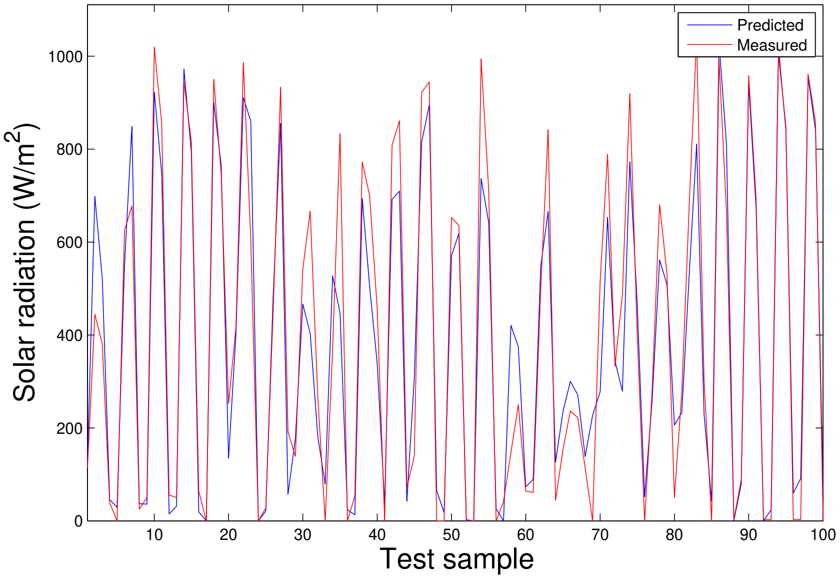


Figure 4.6: Global solar radiation prediction in time by the ELM without feature selection.

Table 4.2: Comparative results of the solar radiation prediction before and after feature selection with the GGA-ELM considering crossovers $\mathcal{C}_1$ and $\mathcal{C}_2$, and dynamics $\mathcal{D}_1$ and $\mathcal{D}_2$.

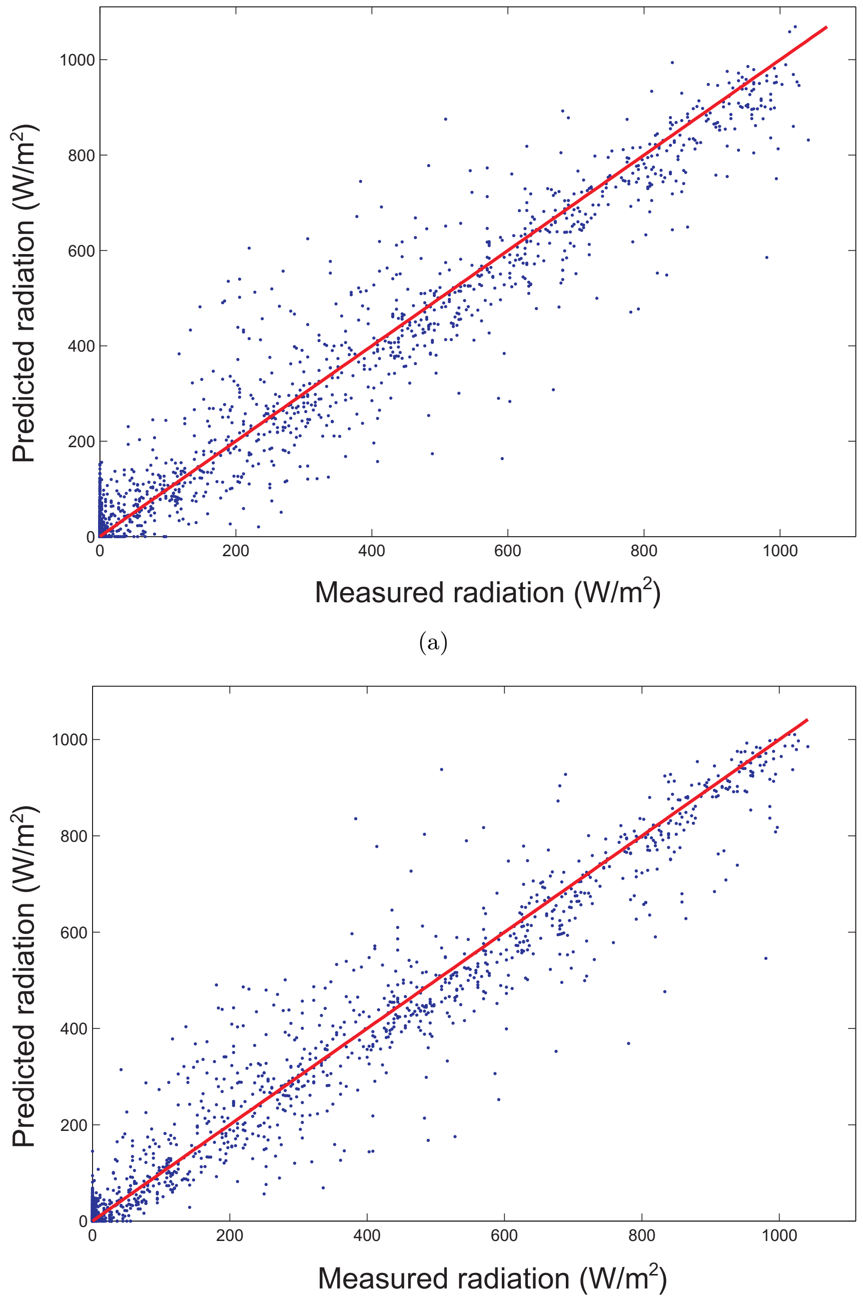
Experiments	RMSE (W/m^2)	r^2
ELM (all features)	85.14	0.9283
GGA-ELM ($\mathcal{C}_1$, $\mathcal{D}_1$)	78.06	0.9382
GGA-ELM ($\mathcal{C}_1$, $\mathcal{D}_2$)	76.14	0.9406
GGA-ELM ($\mathcal{C}_2$, $\mathcal{D}_1$)	77.53	0.9401
GGA-ELM ($\mathcal{C}_2$, $\mathcal{D}_2$)	75.56	0.9415

(two-parents/one-child) and $\mathcal{C}_2$ (two-parents/two-children). It can be observed in Table 4.2 that $\mathcal{C}_1$ presents slightly worse predictions than $\mathcal{C}_2$. Observing the evolution along generations for the first crossover operator ($\mathcal{C}_1$), a continuous increase in the number of groups was detected, and an upper bound in the number of groups to be created was imposed. Therefore, the number of groups was restricted to a maximum of 10, 15, 20 or 25, and any group created over this limit was destroyed and its items were randomly reallocated to existing groups. In spite of this consideration, $\mathcal{C}_2$ was still found to outperform $\mathcal{C}_1$ in this problem. There is a reason detected about this continuous increase in the number of groups when using $\mathcal{C}_1$ crossover: The ELM's output variability and the little variations between different groups causes a large diversity in our search space. At the same time, few features provide valuable information to increase the quality of the measure, then, its easier for the algorithm to get a large number of groups with a low number of features on them.

The last experiments compare the two different dynamics introduced in Section 4.3.3. The first one, $\mathcal{D}_1$, computes the fitness function for all groups in each individual, and the minimum value for all groups is set as the individual's fitness value. The second one, $\mathcal{D}_2$, maximizes each individual's first group fitness function and assigns it as the individual's fitness value. Therefore, $\mathcal{D}_1$ must run the ELMs on all groups in each individual, while $\mathcal{D}_2$ only obtains G_1's fitness function. Which means a large reduction in computation time. Moreover, $\mathcal{D}_1$ always performs the ELMs on the initial number of features, while $\mathcal{D}_2$ removes several features every $\eta = 5$ generations. This applied criterion requires that those features in G_1 that appear at least in 40% of the individuals are removed, and evolution continues with a smaller feature population. Let us focus on the second crossover operator (as it showed the best results), it can be seen that the RMSE decreases from $77.53 W/m^2$, when applying $\mathcal{D}_1$, to 75.56 W/m^2 for $\mathcal{D}_2$, and r^2 increases from 0.9401 to 0.9415, respectively. Figures 4.7 and 4.8 present, respectively, the scatter plot and the solar prediction in time for both dynamics. Once again, it can be seen that the prediction fits rather well to the measured data.

Analyzing the results of the 10-fold cross validation in the best experiment (crossover $\mathcal{C}_2$ and dynamics $\mathcal{D}_2$), it can be determined that the key predictive variables (features) are the OLR, CLDFRA (at a pressure level corresponding to 700 hPa), QVAPOR (at a pressure level corresponding to 700 hPa), and T (at a pressure level corresponding to 500 hPa), all for the WRF output point located at (39° 51'N, 4° 01'W). Other five to nine less important features complete the different GGA-ELM's solutions but their appearance in the individuals of each best solution is not enough to be represented.

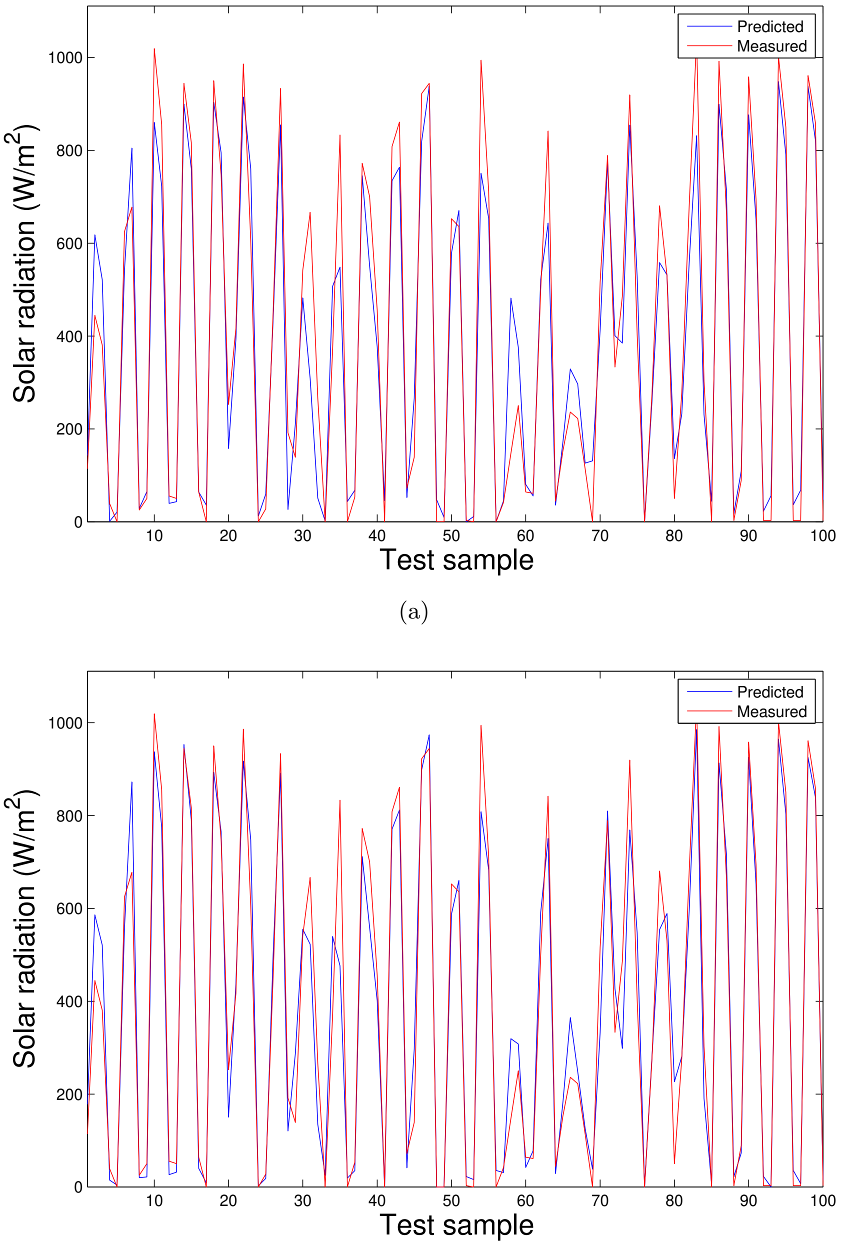
Finally, Figure 4.9 shows the amount of features removed after each η generations, and it can be seen that sometimes no features are removed because their appearance is lower than the threshold marked. Moreover, Figure 4.10 presents the GGA-ELM's performance, where it can be observed that right after several characteristics are removed, the RMSE may increase, but on the long run, better results are obtained.



(a)

(b)

Figure 4.7: Scatter plot of the global solar radiation prediction after feature selection with $\mathcal{C}_2$ crossover operator, and following dynamics: (a) $\mathcal{D}_1$; (b) $\mathcal{D}_2$.



(a)

(b)

Figure 4.8: Global solar radiation prediction in time after feature selection with $\mathcal{C}_2$ crossover operator, and following dynamics: (a) $\mathcal{D}_1$; (b) $\mathcal{D}_2$.

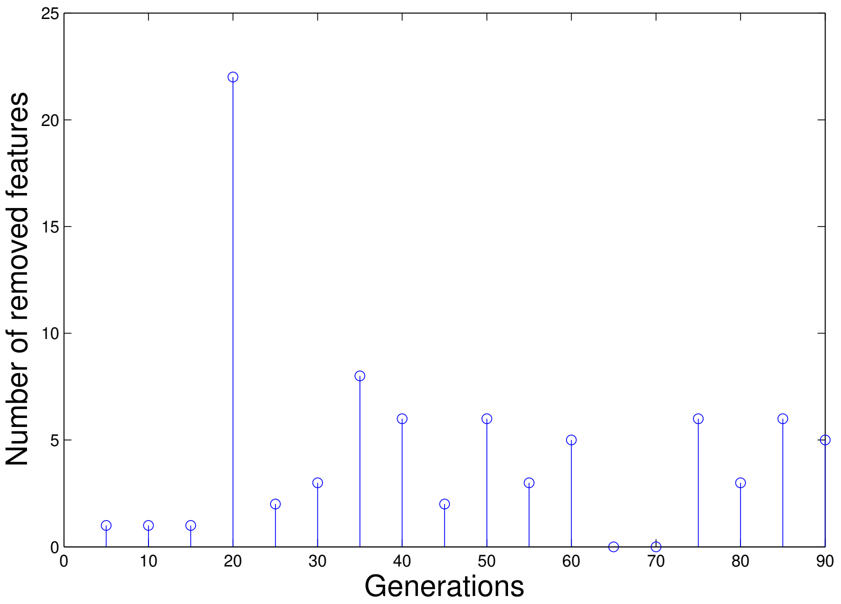


Figure 4.9: Feature removal along generations with $\mathcal{C}_2$ crossover operator, and following dynamics $\mathcal{D}_2$.

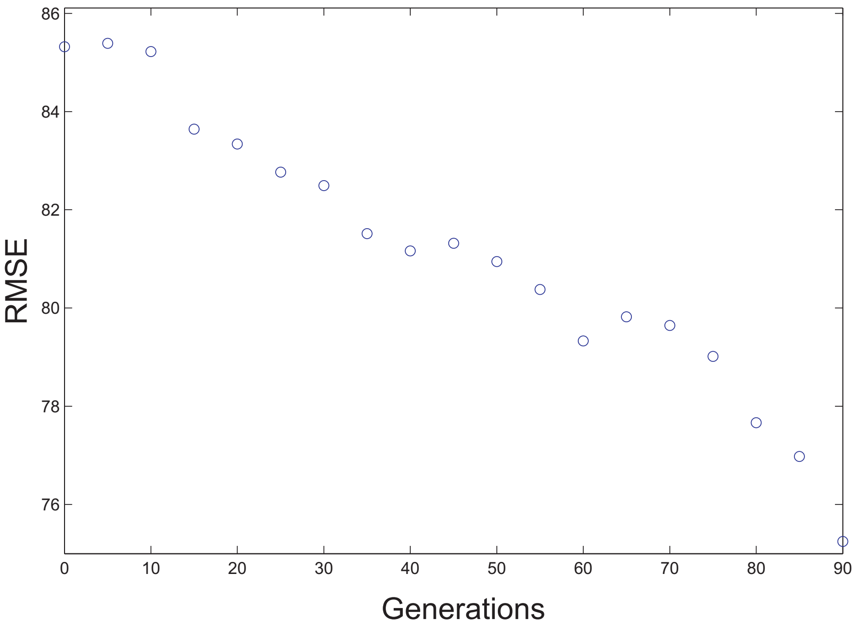


Figure 4.10: GGA-ELM's performance along generations when $\eta = 5$ for $\mathcal{C}_2$ crossover operator, and following dynamics $\mathcal{D}_2$. Note that after η generations, several features are removed.

4.4.3 Subproblem 2

Subproblem 2 analyzes the prediction performance of the proposed GGA-ELM approach for a 1, 2, and 3 hour ahead forecasting, considering only the best algorithms' configuration found in the previous subproblem (i.e. $\mathcal{C}_2$ crossover operator, and dynamics $\mathcal{D}_2$). Note that in this subproblem only the outputs of the WRF model are used as predictive variables. Table 4.3 presents the results obtained for this experiment, as well as the forecasting skill of the proposed GGA-ELM algorithm at point $\mathcal{P}$.

Table 4.3: Global solar radiation prediction for a 1, 2, and 3 hour ahead forecasting, after performing a feature selection with the GGA-ELM (considering crossover $\mathcal{C}_2$ and dynamics $\mathcal{D}_2$).

GGA-ELM ($\mathcal{C}_2$, $\mathcal{D}_2$)	RMSE	r^2	s
Forecast horizon: t+x	(W/m^2)		(ref: $\mathcal{I}_t^{Per}$)
x=1 hour	111.76	0.8693	13%
x=2 hours	165.86	0.7173	30%
x=3 hours	200.36	0.5900	40%

It can be seen that the longer the forecast horizon is, the worse the proposed algorithm predicts the global radiation (i.e. r^2 falls from 0.8693 to 0.5900). On the other hand, the forecast skill shows that the GGA-ELM outperforms the persistence model as the forecast horizon increases (FS from 13% to 40%). This situation is just what we expected during the process, because the information given by the features is lower when the longer the forecast horizon is.

4.4.4 Subproblem 3

Subproblem 3 is also a forecasting experiment that analyzes prediction performance for a time horizon $t + x$, $x = 0, \ldots, \mathcal{X}$. In this subproblem, objective global solar radiation data measured in Toledo's station for times $t - z$, $z = 0, \ldots, \mathcal{Z}$, are included as predictive variables. Note that in this case, as in Subproblem 2, only the best algorithms' configuration (i.e. $\mathcal{C}_2$ crossover operator, and dynamics $\mathcal{D}_2$) and best features found in Subproblem 1 have been analyzed.

Table 4.4 presents the results obtained for a one and two hours ahead prediction, when past and known global solar radiation data are included. Experiments considering station measurements as input variables (ranging from one past value ($z = 0$) to six past values ($z = 0$ to 5)) are shown. For comparison purposes, forecast skill (s) is included.

It can be seen that, by performing a feature selection with the proposed approach, the prediction skill improves over the use of all 92 WRF predictive variables. It is important to highlight that when solar radiation from the previous hours are also considered as input, performance

Table 4.4: Global solar radiation prediction for a 1 and 2 hours ahead forecasting ($\hat{\mathcal{I}}_{t+x}$). Input variables include those features selected with the GGA-ELM approach at Subproblem 1 (crossover $\mathcal{C}_2$ and dynamics $\mathcal{D}_2$), together with past station's measurements ($\mathcal{I}_{t-z}$), ranging from $z = 0$ to $z = 5$.

	Predictive variables: ELM (all features) + $\mathcal{I}_{t-z}$			Predictive variables: GGA-ELM's best features + $\mathcal{I}_{t-z}$		
	RMSE (W/m^2)	r^2	s (ref: $\mathcal{I}^{Per}_{t+x}$)	RMSE (W/m^2)	r^2	s (ref: $\mathcal{I}^{Per}_{t+x}$)
Forecast horizon: x=1						
$z = 0$	106.62	0.8838	17%	100.12	0.8970	22%
$z = 0,\ 1$	101.82	0.8935	21%	82.37	0.9299	36%
$z = 0,\ to\ 2$	95.43	0.9074	26%	78.67	0.9366	39%
$z = 0,\ to\ 3$	92.16	0.9141	28%	78.28	0.9377	39%
$z = 0,\ to\ 4$	90.37	0.9182	30%	76.86	0.9405	40%
$z = 0,\ to\ 5$	90.96	0.9166	29%	76.53	0.9407	41%
Forecast horizon: x=2						
$z = 0$	171.59	0.6921	27%	154.81	0.7317	35%
$z = 0,\ 1$	149.90	0.7652	37%	126.89	0.8311	46%
$z = 0,\ to\ 2$	133.88	0.8147	43%	116.64	0.8585	51%
$z = 0,\ to\ 3$	130.17	0.8264	45%	113.69	0.8670	52%
$z = 0,\ to\ 4$	128.08	0.8321	46%	113.46	0.8678	52%
$z = 0,\ to\ 5$	126.70	0.8340	46%	112.07	0.6899	53%

increases as well. Moreover, increase due to previous data seems to contribute more than any other inputs. In all cases the GGA-ELM's skill is better than the persistence model.

4.5 Conclusions

In this experiment we have presented a novel hybrid Grouping Genetic Algorithm–Extreme Learning Machine (GGA-ELM) approach for accurate global solar radiation prediction problems. The GGA is included to obtain a reduced number of features for the prediction, and the ELM is used as a fast predictor for the solar radiation. The outputs from a numerical weather model (WRF) are used as input features for the ELM, to be selected by the GGA. A real solar radiation prediction problem for Toledo's radiometric observatory (Spain) has been tackled to show the goodness of the proposed approach.

Three subproblems have been analyzed then: First, in Subproblem 1, the prediction system proposed only uses the output of the WRF as inputs, without any other additional information to do the prediction. This case consists of a downscaling of the global solar radiation prediction to a point of interest. In this first subproblem we have introduced and tested different refinements to the GGA-ELM to improve the feature selection and prediction capabilities of the system: 1) two different GGA crossover operators, and 2) two different dynamics for the algorithm, one implying the ELM's error minimization of any group of the GGA, and the second one implying the ELM's error maximization for a specific group of the GGA, followed by the removal and re-initialization of the algorithm afterwards. For this first subproblem, we have found out that the best algorithm's configuration consists of a two-parents/two-children crossover plus the maximization, removal and re-initialization dynamics, which obtains the best results in terms of different error measures. This algorithm's configuration leads to a best solution with only 9 predictive features out of the initial 92.

The second and third subproblems are prediction problem, where we have tried to predict the solar radiation at the point of interest at different time tags $t+x$, but again using predictive variables from the WRF only. In this case, the longer the forecast horizon, the better the GGA-ELM's performance is, in terms of different error measures and forecast skill. Finally, we have tackled a complete prediction problem by including previous values of measured solar radiation (as features for the ELM) plus the predictive variables from the WRF. We have proven that the inclusion of these previous radiation measures significantly improves the forecast skill with respect to a base-line model.

Chapter 5

A CRO-SP Optimization Scheme for Robust Global Solar Radiation Statistical Downscaling

5.1 Introduction

In this chapter we propose a novel hybrid approach formed by a CRO-SP and an ELM to optimally predict the Global Solar Radiation at a given location. We consider again a vast set of meteorological and atmospheric variables provided by a numeric meso-scale model, the WRF, at different points close to the target location under study. Then, a CRO-SP algorithm is used to determine the best subset of WRF variables that lead to a best forecast. In this case the CRO-SP is well-suited for optimization problems with variable-length encodings. In our work, each species in the CRO-SP algorithm represents the use of a different number of WRF variables, which quality will be evaluated by the ELM prediction accuracy.

5.2 Problem considered and input variables

The problem considered in this case is similar to that described in Section 5.2 for Subproblem 1. In this case the WRF model outputs considered in the study provides 116 variables, more than in the previous example, which are the following:

- OLR: Top of atmosphere outgoing long-wave radiation (W/m^2).

- GLW: Downward long-wave flux at ground level (W/m^2).

- SWDOWN: Downward short-wave flux at ground level (W/m^2).

- u: Zonal wind component at different pressure levels (m/s).

- v: Meridional wind component at different pressure levels (m/s).

- w: Vertical wind component at different pressure levels (m/s).

- PSFC: Atmospheric pressure at ground level (hPa).

- QVAPOR: Water vapor mixing ratio (in kg/kg). This variable is defined as the ratio of the mass of water vapor to the mass of dry air.

- TSK: Surface skin temperature (K).

- TH2: Potential temperature at 2 meters above the ground (K).

- T': Perturbation potential temperature (in K) at different pressure levels. The relationship between the perturbation potential temperature, T', and the potential temperature, θ, is $\theta = T' + 300$.

- CLDFRA: Total cloudiness (fraction of clouds in each cell) at different pressure levels. Cloud fraction ranges from 0 (no clouds) to 1 (clouds in a spatial grid cell).

Table 4.1 shows the 58 variables analyzed for each grid point considered, indicating (when needed) the different pressure levels where they were obtained. Therefore, as two grid points ($M = 2$) have been examined, a total of 116 variables have been used in this work.

5.3 Methodology

We use now the CRO-SP (Section 3.3.1) to determine which set of WRF outputs obtains the best global solar radiation prediction. As in the previous case, fitness function considered for each coral (individual) is obtained computing the Root Mean Square Error (RMSE) of the global solar radiation prediction given by the ELM, as shown in Equation (5.1), where, again, $\mathcal{I}_t$ stands for the global solar radiation measured at a time t, $\hat{\mathcal{I}}_t$ stands for the global solar radiation estimated by the ELM, and T stands for the number of samples in the test set.

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\hat{\mathcal{I}}_t - \mathcal{I}_t\right)^2} \tag{5.1}$$

5.4 Experiments and results

This section describes the experiments run in this problem of solar radiation estimation. Again, the aim is to predict the global solar radiation at a point P using as predictive variables the outputs of the WRF model obtained in two grid points close to P. The first step is to determine the best predictive variables (feature selection) and has been addressed implementing a CRO-SP algorithm with the parameters shown in Table 5.1.

Table 5.1: CRO optimization parameters.

Phase	Parameter
Inicialization	Reef size $= 50 \times 40$ (2,000 positions) $\mathcal{S}_i$, $i \in \{1..5\}$ (5 species) $\rho_0 = 0.75$ (1,500 corals) $\rho_0^{\mathcal{S}_i} = 0.15$ (300 corals per species)
External sexual reproduction	$F_b = 0.70$ Random selection of broadcast spawners. Each possible coral must be broadcast spawner at least once per iteration k. New larva formation using 2-point crossover.
Internal sexual reproduction	$1 - Fb = 0.20$ $P_i = 0.30$
Larvae setting	$\eta = 3$ Identical corals are not allowed in the reef.
Asexual reproduction	$F_a = 0.05$ $P_a = 0.005$
Depredation	$F_d = 0.15$ $P_d = 0.25$ (it decreases with the number of iterations. At k_{max}, $P_d = 0$)
Stop criteria	$k_{max} = 300$ iterations.

To identify the best set and number of predictive variables, several experiments ($\mathcal{E}_i$, $i \in [1,\ldots,3]$) have been run in a 10-fold cross validation scheme. Each CRO-SP experiment $\mathcal{E}_i$ consists of five subexperiments, each one of them analyzing a specific species $\mathcal{S}_i$ ($i \in [1,\ldots,5]$), i.e., all corals belonging to one species have the same number of features. The co-evolution of these species leads quickly to a coral-reef colonized by the most suited corals. Once convergence is reached, the best coral in the reef belongs to a specific species and its health function stands for the RMSE value obtained in test. Note that to calculate the global solar radiation at each iteration, the ELM has been run 3 times and the health function value assigned to the coral is the average result obtained.

Table 5.2 presents the results obtained for each experiment: the average value of error in test, the best coral's RMSE and its corresponding species (in terms of the number of predictive variables in that species).

The first experiment run, $\mathcal{E}_1$, is meant to resolve the order of magnitude of the number of features to be considered (10, 20, 30, 40 or 50), and it can be observed that the best prediction is found using 10 variables (RMSE = 68.21 W/m^2). Experiments $\mathcal{E}_2$ and $\mathcal{E}_3$ are used to refine

Table 5.2: Experiments run considering different species. Each *species* is represented by $\mathcal{S}_i$

Experiment	Number of					RMSE		Best Species
	features per species					(W/m^2)		
	$\mathcal{S}_1$	$\mathcal{S}_2$	$\mathcal{S}_3$	$\mathcal{S}_4$	$\mathcal{S}_5$	Average	Best coral	
$\mathcal{E}_1$	10	20	30	40	50	69.50	68.21	10 features ($\mathcal{S}_1$)
$\mathcal{E}_2$	6	8	10	12	14	69.33	68.16	8 features ($\mathcal{S}_2$)
$\mathcal{E}_3$	7	8	9	10	11	69.16	68.03	8 features ($\mathcal{S}_2$)

the number of predictive variables to consider, both of them converging to best results when the species encode 8 variables. Figure 5.1 presents the scatter plots for each experiments' best coral, showing the algorithm's good performance in all cases.

Figure 5.2 presents the comparison in time between the measured and the predicted GSR for experiment $\mathcal{E}_3$, the best experiment, where it can be seen that the prediction follows rather well the field (target) data. Figure 5.3 shows the evolution with the number of iterations of the best coral in this experiment. It corresponds to a coral encoding the use of 8 WRF variables and presents a final RMSE of 68.03 W/m^2 and a coefficient of determination r^2 =0.9531.

It is interesting to analyze the behavior (evolution) of the different species in the reef with the number of iterations. Figure 5.4 shows this evolution for the best run of the best experiment, $\mathcal{E}_3$. In Figure 5.4(a) the random initialization of the reef is presented, where the reader can see the positions occupied by each different species and the free positions available at the reef. As the number of iterations (k) increases (Figures 5.4(b)-(f)), it can be observed that the worst-fitted species tend to die and are no longer present at the reef, as larvae from dominant species outperform them. Finally, when the stop criteria is reached, the reef is colonized by the best species which, in this particular experiment, is species $\mathcal{S}_2$ (corresponding to the use of 8 WRF variables for the prediction).

Figures 5.5 to 5.7 show, for all experiments analyzed, the evolution with the number of iterations of two important characteristics. First, the root mean square error of each species' best coral, which is depicted in subfigures (a). It is clear that the RMSE decreases with the number of evolutions, but there is one exception: when a species is endangered (is being outperformed by the rest) its RMSE increases abruptly. Right after this occurs, the RMSE is interrupted, resulting in the disappearance of the worst-fitted species from the reef. Second, the number of corals present in each species is analyzed in Figures 5.5(b), 5.6(b) and 5.7(b). It can be observed that, at some points, the number of corals in some species drops down. This is directly related to the occurrence of depredation phases. It is important to highlight that although in the depredation phase the species are decimated, the evolution keeps recovering the best-fitted.

Next, Table 5.3 shows the name of the best coral's WRF outputs selected for each experiment. It can be seen that there are six variables: OLR, w at 400 hPa, CLDFRA at 200 hPa, T at 850 hPa and T at 400 hPa corresponding to the first grid point, and v at 500 hPa corresponding to the second grid point, present in all experiments' results. Therefore, we can conclude that these variables set the rough prediction while the other WRF outputs selected by the algorithm perform the refinement. Thus, for the third experiment, the RMSE using these 6 variables (over the same test sets) is 74.05 W/m^2 and 72.59 W/m^2, average and best values respectively. Once the refinement takes place, these RMSE values drop down to 69.16 W/m^2 and 68.03 W/m^2 respectively (as stated in Table 5.2).

Table 5.3: Best predictive variables found for each experiment. Those variables present in all three experiments' results have been highlighted in bold face.

Experiment	Best WRF outputs selected	
	Grid point #1	Grid point #2
$\mathcal{E}_1$	**OLR**, ***w* (400 hPa)**, **CLDFRA (200 hPa)**, **T (850 hPa)**, **T (400 hPa)**, u (100 hPa)	***v* (500 hPa)**, PSFC, TH2, u (50 hPa)
$\mathcal{E}_2$	**OLR**, ***w* (400 hPa)**, **CLDFRA (200 hPa)**, **T (850 hPa)**, **T (400 hPa)**,	***v* (500 hPa)**, TH2 w (300 hPa)
$\mathcal{E}_3$	**OLR**, ***w* (400 hPa)**, **CLDFRA (200 hPa)**, **T (850 hPa)**, **T (400 hPa)**,	***v* (500 hPa)**, u (850 hPa) u (50 hPa)

Finally, in Table 5.4 the results are compared to those obtained with other techniques. First, the reader can see the GSR prediction using the 116 WRF variables (no feature selection) as inputs to the ELM. Then, feature selection is performed using three different techniques: a Genetic Algorithm (GA), a GGA (as described in [6]) and the proposed CRO-SP approach, and the variables chosen are used as the inputs to the ELM. It can be seen that the best results are obtained when the CRO with species is used.

5.5 Conclusions

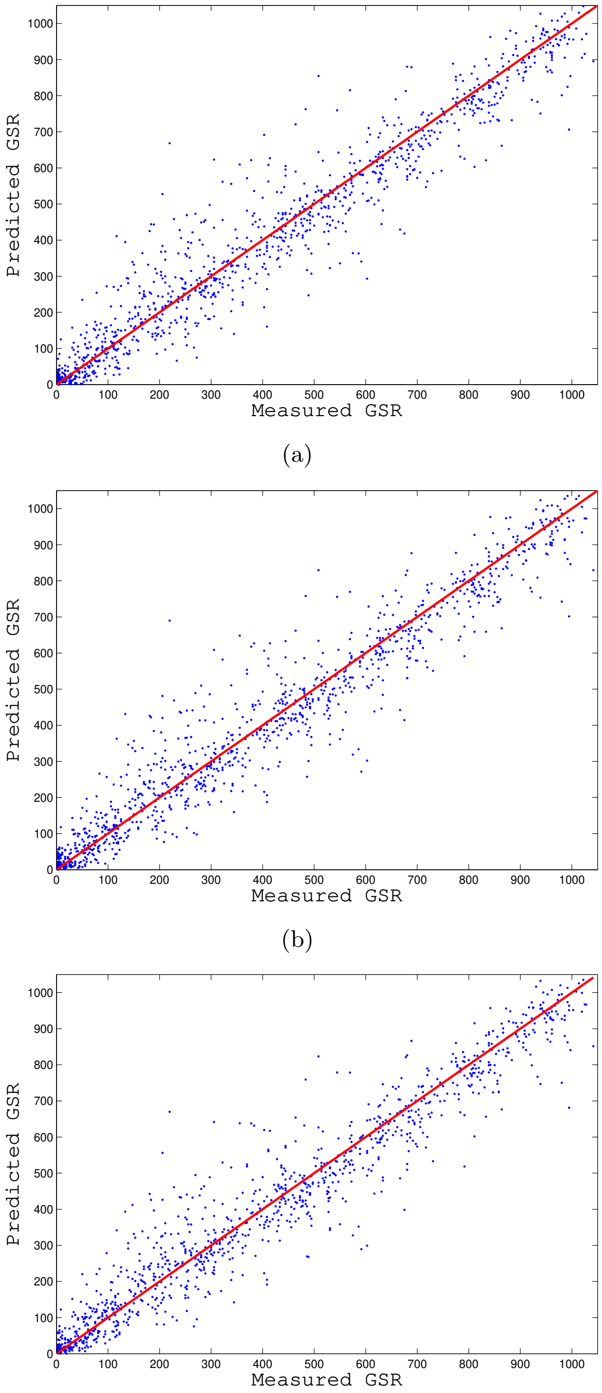
In this chapter we have tackled a global solar radiation prediction problem by using a novel co-evolution algorithm *Coral Reefs Optimization algorithm with species* (CRO-SP), combined with an Extreme Learning Machine (ELM). The ultimate goal of this experiment has been to evaluate what predictive variables from the numerical weather model (i.e. the WRF model)

Table 5.4: Comparison of the results obtained with other metaheuristic techniques.

Metaheuristic technique	RMSE (W/m^2)	
	Average	Best individual
No feature selection + ELM	88.24	87.25
Genetic algorithm + ELM	73.98	72.20
Grouping Genetic Algorithm + ELM [6]	74.73	73.66
CRO-SP + ELM	69.16	68.21

perform best. For this purpose, each species in the CRO-SP encodes a fixed and different number of variables to be analyzed and the best species comes out as a result of the co-evolution.

To determine the best set and number of predictive variables, three experiments have been run in a 10-fold cross validation scheme and the RMSE has been used as common measure. The experiment where 7, 8, 9, 10 and 11 variables are co-evolved (experiment $\mathcal{E}_3$), produces an average best result of 69.19 W/m^2 an a best result of 68.03 W/m^2, turning in a 21.62 % and 22.03 % improvement, respectively, over the average and best prediction without feature selection.

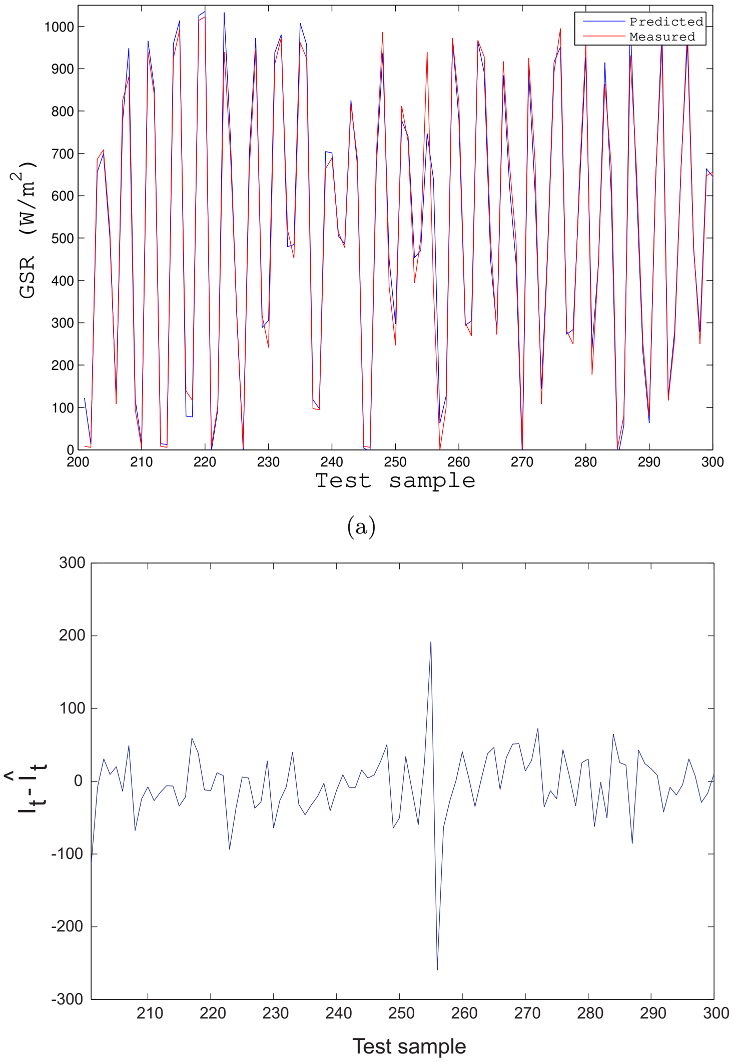


(a)

(b)

(c)

Figure 5.1: Scatter plot of the global solar radiation: (a) Experiment $\mathcal{E}_1$, (b) Experiment $\mathcal{E}_2$ and (c) Experiment $\mathcal{E}_3$.



(a)

(b)

Figure 5.2: Experiment $\mathcal{E}_3$. (a) Global solar radiation in time. (b) Deviation in time of the predicted GSR from the measured GSR. Note that only a random time frame of 100 samples is presented for clarity purposes.

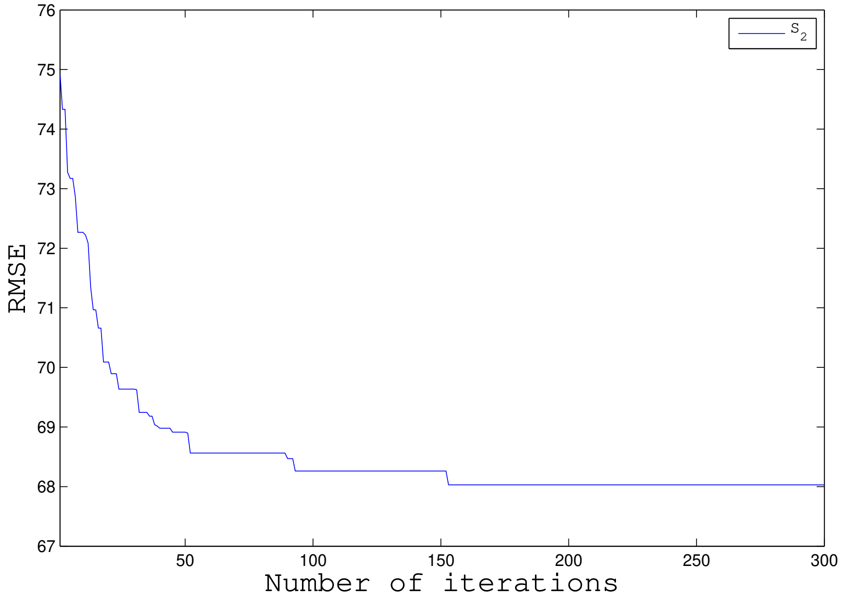
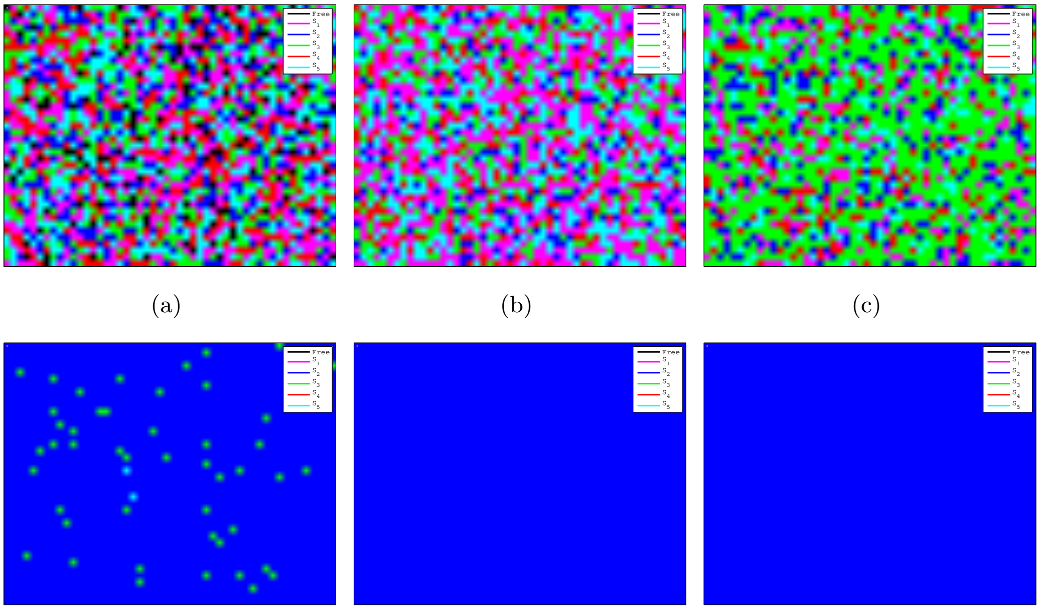


Figure 5.3: Experiment $\mathcal{E}_3$. Evolution with the number of iterations of the best coral's RMSE. Note that the best coral belongs to species $\mathcal{S}_2$.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 5.4: Experiment $\mathcal{E}_3$. Evolution of the species present in the reef after a certain number of iterations (k): (a) $k = 1$, (b) $k = 10$, (c) $k = 25$, (d) $k = 50$, (e) $k = 150$, (f) $k = 300$.

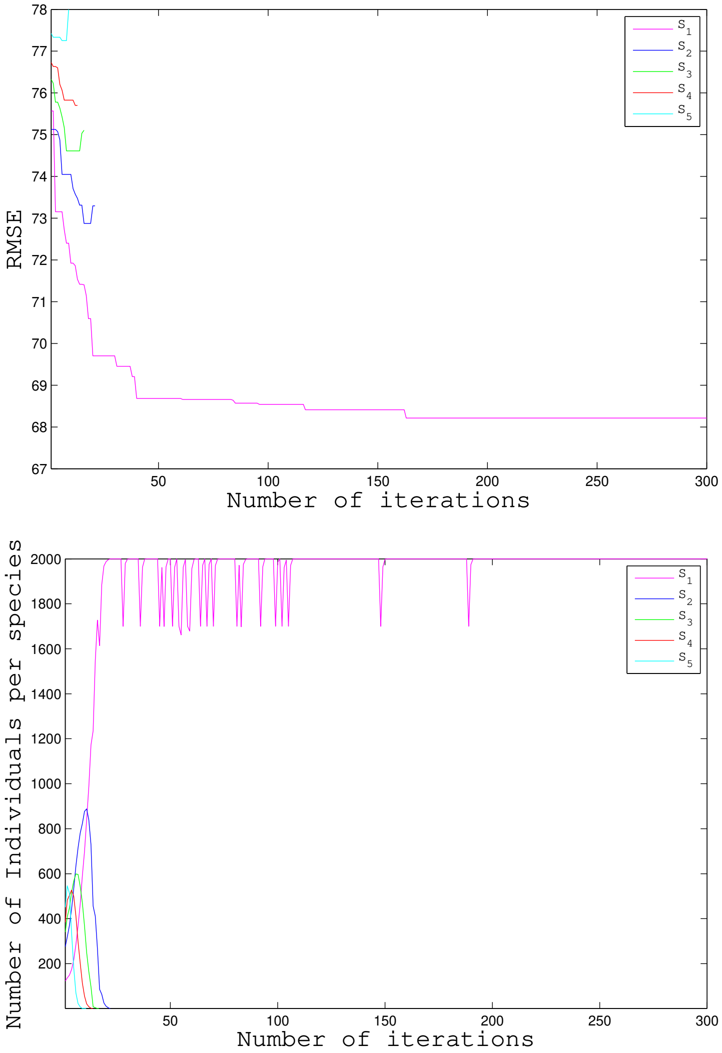


Figure 5.5: Experiment $\mathcal{E}_1$. Evolution with the number of iterations of: (a) RMSE of each species' best coral, and (b) Number of corals per species.

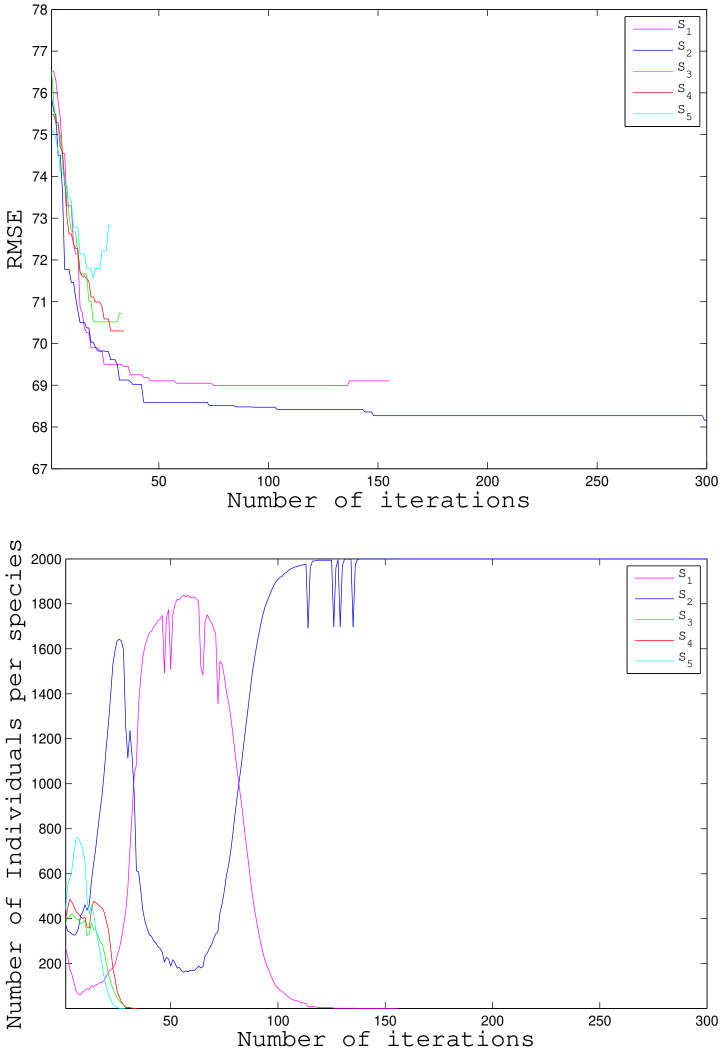


Figure 5.6: Experiment $\mathcal{E}_2$. Evolution with the number of iterations of: (a) RMSE of each species' best coral, and (b) Number of corals per species.

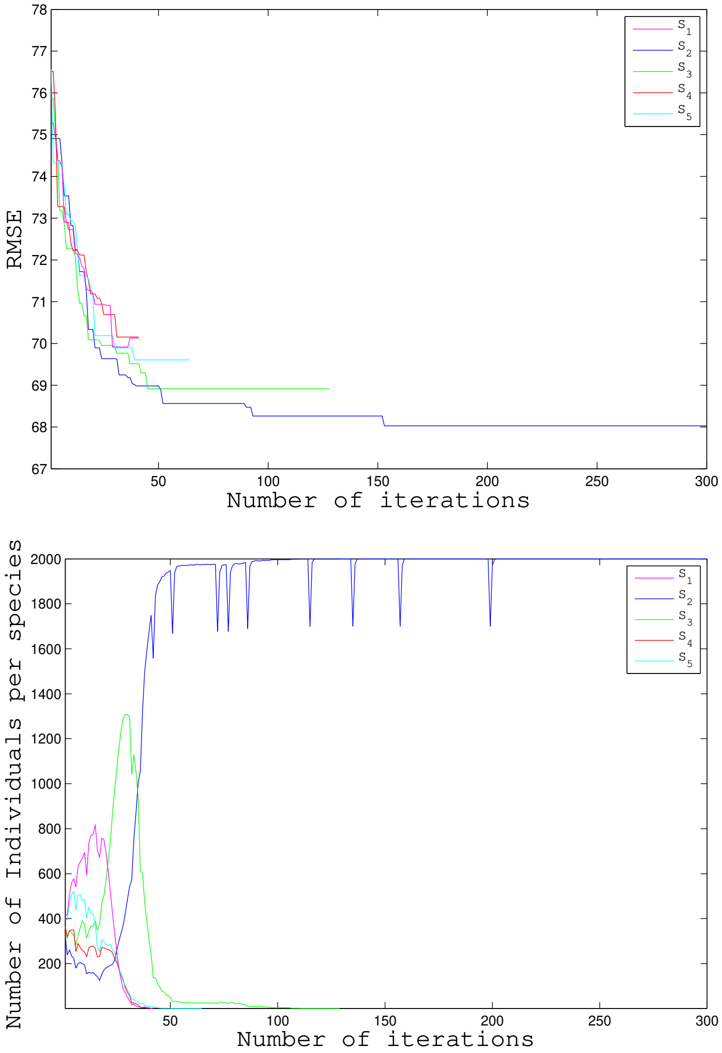


Figure 5.7: Experiment $\mathcal{E}_3$. Evolution with the number of iterations of: (a) RMSE of each species' best coral, and (b) Number of corals per species.

Part IV

Thesis conclusions and future research

Chapter 6

Conclusions and Future Work

6.1 Conclusions

This Ph.D. thesis deals with the problem of global solar radiation prediction at a given point from numerical weather models, specifically the WRF meso-scale model. We propose hybrid neuro-evolutionary algorithms to deal with the large amount of features from the WRF, by means of wrapper feature selection. For this, we propose the GGA-ELM approach in Chapter 4 and a hybrid CRO-SP plus ELM approach has been proposed in Chapter 5.

The use of this kind of techniques, based on hybrid techniques related to bio-inspired algorithms [25], (either traditional such as evolutionary computation techniques, or new approaches such as the Coral Reefs Optimization algorithm) produces strong and robust techniques for solar radiation prediction, in which feature selection methods are integrated together with the prediction algorithms, carried out by a fast-training neural network.

From the results of this research, different conclusions can be raise, summarized here:

1. Hybrid prediction systems for optimal global solar radiation prediction have been proposed in this work, where the WRF provides a prediction of atmospheric variables at different pressure levels at a given point, that will be used as inputs in the prediction system (ELM) to estimate the global solar radiation prediction, and improved (by means of feature selection) with different hybrid bio-inspired algorithms.

2. A hybrid GGA-ELM has been proposed to obtain an optimal prediction of solar global radiation at a given point of the Earth surface. The GGA algorithm proposed obtains the best features from the WRF model to do the prediction, while the ELM will carry out the final prediction of the global solar radiation. This is the first time that a grouping-based algorithm is proposed for feature selection. Two different dynamics and crossovers are proposed in the GGA: D1 is a regular dynamic proposed in a GGA, in which all the groups of features are evaluated by means of the ELM, and the best group is kept as the solution to the feature selection. On the other hand, D2 is a novel contribution of

this These, in which groups of the worst features from the WRF model are selected by the GGA, and erased from the evolution to achieve the best prediction with the survival features. This D2 has proven to be less computational costly than D1, and while obtaining even better solutions. Regarding the crossover operators, C1 is a regular crossover operator and C2 is a double crossover which provides extra diversity in the solutions, leading to better results.

3. A hybrid CRO-SP and ELM system has been proposed to optimally predict the Global Solar Radiation at a given location. In this case, the hybrid approach is able to cope with different encoding for individuals, dealing this way with individuals that encoded from 6 to 50 features, and co-evolved them to obtain the best solar radiation prediction. The contribution of this system is the application of a novel co-evolution algorithm (CRO-SP+ELM) to predict the global solar radiation and improved the results previously obtained with other techniques.

6.2 Future Research

Despite of the different results and lines obtained from this Ph.D. thesis, there are several directions in which studies could develop new research. Some of the research works that could be addressed in the near future are following:

1. This thesis focuses in a meso-scale model, the WRF, as input generator for the different hybrid models applied. Next research could be the evaluation of the potential of the different hybrid models proposed with alternative NWM, or combination of them, to generate new input variables to the hybrid system.

2. Different meteorological, energy and renewable energy prediction and optimization problems could be proposed, such as [96] and [23], among others, to apply both algorithms evaluated in this thesis and developing new results and improvements.

3. The ELM is applied as a regressor system to obtain the best prediction from the features selected. Other regressors with enough fast training could be applied to hybridize with the bio-inspired feature selection algorithms proposed, evaluating computational load, execution timings and number of iterations to achieve a good solution within a reasonable computational time.

4. This thesis focuses on mono-objective optimization problems. A strong extension of this research would be to apply and adapt CRO the algorithm [99], both species evolution and substrate layer 3, to be applied in multi-objective problems. In the research just proposed, corals and larvaes are under modification to compete for a position in the coral, coevolving with different individuals from different operators, such as genetic operator, bioinspired

and others. In this case, a multi-objective algorithm obtains a Pareto front with different number of best solutions on the contrary as expected in a mono-objective problem.

Part V

Appendix

Appendix A. List of publications

This section presents a summary of the scientific publications obtained during the research carried out in this Ph.D. Thesis, in addition to the publications of the Compendium. As can be seen, the research work carried out has been extensive and has covered very different algorithms in Machine Learning, and also different applications, in addition to solar radiation prediction.

6.3 Papers in International Journals

1. S. Salcedo-Sanz, R. García-Herrera, C. Camacho-Gómez, **A. Aybar Ruiz** and E. Alexandre, "Wind power field reconstruction from a reduced set of representative measuring points," *Applied Energy*, vol. 228, pp. 1111-1121, 2018. (JCR: 7.900, Q1)

2. S. Salcedo-Sanz, **A. Aybar-Ruiz**, C. Camacho-Gómez and E. Pereira, "Efficient fractal-based mutation in evolutionary algorithms from iterated function systems," *Communications in Nonlinear Science and Numerical Simulation*, vol. 56, pp. 434-446, 2018. (JCR: 2.784, Q2)

3. L. Cornejo-Bueno, C. Camacho-Gómez, **A. Aybar-Ruiz**, L. Prieto, A. Barea-Ropero and S. Salcedo-Sanz,"Wind power ramp event detection with a hybrid neuro-evolutionary approach," *Neural Computing and Applications*, vol. 32, no. 2, pp. 391-402, 2020. (JCR: 4.774, Q1)

6.4 Papers in International Conferences

1. L. Cornejo-Bueno, **A. Aybar Ruiz**, S. Jiménez-Fernández, E. Alexandre, J. C. Nieto-Borge and S. Salcedo-Sanz, "A grouping genetic algorithm â€" Extreme learning machine approach for optimal wave energy prediction," 2016 IEEE Congress on Evolutionary Computation (CEC), Vancouver, BC, pp. 3817-3823, 2016.

2. L. Cornejo-Bueno, **A. Aybar-Ruiz**, C. Camacho-Gómez, L. Prieto, A. Barea-Ropero and S. Salcedo-Sanz, "A Hybrid Neuro-Evolutionary Algorithm for Wind Power Ramp Events Detection," IWANN, Cádiz, Spain, pp. 745-756, 2017.

3. L. Cornejo-Bueno, C. Camacho-Gómez, **A. Aybar-Ruiz** , L. Prieto, and S. Salcedo-Sanz, “Feature Selection with a Grouping Genetic Algorithm – Extreme Learning Machine Approach for Wind Power Prediction,” CAEPIA, Conference of the Spanish Association for Artificial Intelligence, Salamanca, Spain, pp. 373-382, 2016.

4. **A. Aybar-Ruiz**, J. Del Ser, J.A. Portilla-Figueras and S. Salcedo-Sanz,“A Grouping Harmony Search Algorithm for Assigning Resources to Users in WCDMA Mobile Networks,” ICHSA, International Conference on Harmony Search Algorithm, Bilbao, Spain, pp. 190-199, 2017.

6.5 National conferences

1. L. Cornejo-Bueno, C. Camacho-Gómez,**A. Aybar-Ruiz**, L. Prieto and S. Salcedo-Sanz, “Feature Selection with a Grouping Genetic Algorithm - Extreme Learning Machine Approach for Wind Power Prediction”, XI Congreso Español de Meta-heurísticas, Algoritmos Evolutivos y Bioinspirados (MAEB 2016), Salamanca, España, pp. 373-382, 2016.

Bibliography

[1] S. Agrawal, S. Panda and A. Abraham, “A Novel Diagonal Class Entropy-Based Multilevel Image Thresholding Using Coral Reef Optimization,” *IEEE Transactions on Systems, Man and Cybernetics*, pp. 1-9, 2018.

[2] M. A. Alharbi, “Daily global solar radiation forecasting using ANN and Extreme Learning Machines: a case study in Saudi Arabia,” Master of Applied Science Thesis, Dalhousie University, Halifax, Nova Scotia, 2013.

[3] M. Alizamir, S. Kim, O. Kisi and M. Zounemat-Kermani, “A comparative study of several machine learning based non-linear regression methods in estimating solar radiation: Case studies of the USA and Turkey regions,” *Energy*, vol. 197, no. 117239, 2020.

[4] F. Almonacid, E. F. Fernandez, A. Mellit and S. Kalogirou, “Review of techniques based on artificial neural networks for the electrical characterization of concentrator photovoltaic technology,” *Renewable and Sustainable Energy Reviews*, vol. 75, pp. 938-953, 2017.

[5] A. Angstrom, “Solar and terrestrial radiation. Report to the international commission for solar research on actinometric investigations of solar and atmospheric radiation,” *Quarterly Journla of the Royal Meteorology Society*, vol. 50, pp. 121-126, 1924.

[6] A. Aybar Ruiz, A., S. Jiménez-Fernández, L. Cornejo Bueno, C. Casanova, J. Sanz-Justo, P. Salvador-González and S. Salcedo-Sanz, “A novel Grouping Genetic Algorithmâ€“Extreme Learning Machine approach for global solar radiation prediction from numerical weather models inputs,” *Solar Energy*, vol. 132. pp. 129-142, 2016.

[7] M. A. Behrang, E. Assareh, A. Ghanbarzadeh and A.R. Noghrehabadi, “The potential of different artificial neural network (ANN) techniques in daily global solar radiation modeling based on meteorological data,”*Solar Energy*, vol. 84, no. 8, pp. 1468-1480, 2010.

[8] M. Benghanem and A. Mellit, “Radial Basis Function Network-based prediction of global solar radiation data: Application for sizing of a stand-alone photovoltaic system at Al-Madinah, Saudi Arabia,” *Energy*, vol. 35, no. 9, pp. 3751-3762, 2010.

[9] E. Bermejo, M. Chica, S. Damas, S. Salcedo-Sanz and O. Cordón, “Coral Reef Optimization with substrate layers for medical image registration,” *Swarm and Evolutionary Computation*, vol. 42, pp. 138-159, 2018.

[10] H.G. Beyer, J. Polo Martinez and M. Suri, “Report on Benchmarking of Radiation Products. Report under contract no. 038665 of MESoR,” <http://www.mesor.net/deliverables. html>. Deliverable 1.1.3. 2011.

[11] S. Bhardwaj, V. Sharma, S. Srivastava, O. S. Sastry, B. Bandyopadhyay, S. S. Chandel and J. R. Gupta, “Estimation of solar radiation using a combination of Hidden Markov Model and generalized Fuzzy model,” *Solar Energy*, vol. 93, pp. 43-54, 2013.

[12] M. Bilgili and M. Ozoren, “Daily total global solar radiation modeling from several meteorological data,” *Meteorology and Atmospheric Physics*, vol. 112, no. 3-4, pp. 125-138, 2011.

[13] F. Birol, “Key World Energy Statistics 2017,” *International Energy Agency (JEA)*, 2017.

[14] C. M. Bishop, *Neural networks for pattern recognition*, Oxford University Press, 1995.

[15] A. Blum and P. Langley, “Selection of relevant features and examples in Machine Learning,” *Artificial Intelligence*, vol. 97, pp. 245-271, 1997.

[16] M. J. Vermeij, “Substrate composition and adult distribution determine recruitment patterns in a Caribbean brooding coral”, *Marine Ecology Progress Series*, Vol. 295, pp. 123-133, 2005.

[17] L. Bottou, “Stochastic gradient learning in neural networks,” *Proceedings of Neuro Nimes*, vol. 91, no. 8 pp. 12, 1991.

[18] C. Camacho-Gómez, X. Wang, E. Pereira, I. M. Díaz and S. Salcedo-Sanz, “Active vibration control design using the Coral Reefs Optimization with Substrate Layer algorithm,” *Engineering Structures*, vol. 157, pp. 14-26, 2018.

[19] C. Camacho-Gómez, I. Marsá, J. M. Giménez-Guzmán and S. Salcedo-Sanz, “A Coral Reefs Optimization algorithm with substrate layer for robust Wi-Fi channel assignment,” *Soft Computing*, in press, 2019.

[20] D. Carvalho, A. Rocha, M. Gómez-Gesteira and C. Silva Santos, “Sensitivity of the WRF model wind simulation and wind energy production estimates to planetary boundary layer parameterizations for onshore and offshore areas in the Iberian Peninsula,” *Applied Energy*, vol. 135, pp. 234-246, 2014.

[21] J. L. Chen, H. B. Liu, W. Wu and D. T. Xie, “Estimation of monthly solar radiation from measured temperatures using support vector machines - A case study,” *Renewable Energy*, vol. 36, pp. 413-420, 2011.

[22] C. Coimbra, J. Kleissl and R. Marquez, “Overview of solar forecasting methods and a metric for accuracy evaluation,” *Solar Resource Assessment and Forecasting*, edited by J. Kleissl, Elsevier, Waltham, Massachusetts, 2013.

[23] L. Cornejo-Bueno, C. Casanova-Mateo, J. Sanz-Justo and S. Salcedo-Sanz, “Machine learning regressors for solar radiation estimation from satellite data,” *Solar Energy*, vol. 183, pp. 768-775, 2019.

[24] K. Dahmani, R. Dizene, G. Notton, C. Paoli, C. Voyant and M. L. Nivet, “Estimation of 5-min time-step data of tilted solar global irradiation using ANN (Artificial Neural Network) model,” *Energy*, vol. 70, pp. 374-381, 2014.

[25] J. Del Ser, E. Osaba, E. Molina, X.S. Yang, S Salcedo-Sanz, D. Camacho et al., “Bio-inspired computation: Where we stand and what’s next,” *Swarm and Evolutionary Computation*, vol. 48, pp. 220-250, 2019.

[26] R. C. Deo and M. Sahin, “Forecasting long-term global solar radiation with an ANN algorithm coupled with satellite-derived (MODIS) land surface temperature (LST) for regional locations in Queensland,” *Renewable and Sustainable Energy Reviews*, vol. 72, pp. 828-848, 2017.

[27] R. C. Deo, M. Sahin, J. F. Adamowski and J. Mi, “Universally deployable extreme learning machines integrated with remotely sensed MODIS satellite predictors over Australia to forecast global solar radiation: A new approach,” *Renewable and Sustainable Energy Reviews*, vol. 104, pp. 235-261, 2019.

[28] M. Diagne, M. David and J. Boland, “Post-processing of solar irradiance forecasts from WRF model at Reunion Island,” *Solar Energy*, vol.105, pp. 99-108, 2014.

[29] H. Dong, L. Yang, S. Zhang and Y. Li,“Improved prediction approach on solar irradiance of photovoltaic Power Station,”*TELKOMNIKA Indonesian Journal of Electrical Engineering*, vol. 12, no.3, pp. 1720-1726, 2014.

[30] A. S. Dorvlo, J. A. Jervase and A. Al-Lawati, “Solar radiation estimation using artificial neural networks,”*Applied Energy*, vol. 71, no. 4, pp. 307-319, 2002.

[31] E. Falkenauer, “The grouping genetic algorithmâ€“widening the scope of the GAs,” *In Proc. of the Belgian journal of operations research, statistics and computer science*, vol. 33, pp. 79-102, 1992.

[32] E. Falkenauer, “Genetic algorithms for grouping problems,” *New York: Wiley*, 1998.

[33] Y. Feng, W. Hao and H. Li, N. Cui, D. Gong and L. Gao, “Machine learning models to quantify and map daily global solar radiation and photovoltaic power,” *Renewable and Sustainable Energy Reviews*, vol. 118, no. 109393, 2020.

[34] A. J. Ferreira and M. T. Figueiredo, “Incremental filter and wrapper approaches for feature discretization,” *Neurocomputing*, vol. 123, pp. 60-74, 2014.

[35] C.-L. Fu and H.-Y. Cheng, “Predicting solar irradiance with all-sky image features via regression,” *Solar Energy*, vol. 97, pp. 537-550, 2013.

[36] L. Garcia-Hernandez, L. Salas-Morera, C. Carmona-Muñoz, J. A. Garcia-Hernandez, S. Salcedo-Sanz, “A novel Island Model based on Coral Reefs Optimization algorithm for solving the unequal area facility layout problem,” *Engineering Applications of Artificial Intelligence*, vol. 89, no. 103445, 2020.

[37] S. Ghimire, R. C. Deo, N. Raj and J. Mi, “Wavelet-based 3-phase hybrid SVR model trained with satellite-derived predictors, particle swarm optimization and maximum overlap discrete wavelet transform for solar radiation prediction,” *Renewable and Sustainable Energy Reviews*, vol. 113, no. 109247, 2019.

[38] S. Ghimire, R. C. Deo, N. J. Downs and N. Raj, “Global solar radiation prediction by ANN integrated with European Centre for medium range weather forecast fields in solar rich cites of Queensland Australia,” *Journal of Cleaning Production*, vol. 216, pp. 288-310, 2019.

[39] S. Ghimire, R. C. Deo, N. Raj and J. Mi, “Deep solar radiation forecasting with convolutional neural network and long short-term memory network algorithms,” *Applied Energy*, vol. 253, no. 113541, 2019.

[40] T.M. Giannaros, V. Kotroni and K. Lagouvardos, “Predicting lightning activity in Greece with the weather research and forecasting (WRF) model,” *Atmospheric Research*, vol. 156, pp. 1-13, 2015.

[41] R. Hashim, C. Roy, S. Motamedi, S. Shamshirband and D. Petkovic, “Selection of climatic parameters affecting wave height prediction using an enhanced Takagi-Sugeno-based fuzzy methodology,” *Renewable and Sustainable Energy Reviews*, vol. 60, pp. 246-257, 2016.

[42] S. Haykin, *Neural networks: a comprenhensive foundation*, Prentice Hall, 1998.

[43] F. O. Hocaoglu, O. N. Gerek and M. Kurban, “Hourly solar radiation forecasting using optimal coefficient 2-D linear filters and feed-forward neural networks,”*Solar Energy*, vol. 82, pp. 714-726, 2008.

[44] G.B. Huang, Q.Y. Zhu and C.K. Siew, “Extreme learning machine: theory and applications,” *Neurocomputing*, vol. 70, no. 1, pp. 489-501, 2006.

[45] S. Huda, M. Abdollahian, M. Mammadov, J. Yearwood, S. Ahmed and I. Sultan, “A hybrid wrapper-filter approach to detect the source(s) of out-of-control signals in multivariate manufacturing process,” *European Journal of Operational Research*, vol. 237, pp. 857-870, 2014.

[46] S. Huda, J. Abawajy, M. Alazab, M. Abdollalihian, R. Islam and J. Yearwood, "Hybrids of support vector machine wrapper and filter based framework for malware detection," *Future Generation Computer Systems*, vol. 55, pp. 376-390, 2016.

[47] J. Huertas-Tato, R. Aler, I. M. Galván, F. J. Rodríguez-Benítez, C. Arbizu-Barrena and D. Pozo-Vázquez, "A short-term solar radiation forecasting system for the Iberian Peninsula. Part 2: Model blending approaches based on machine learning," *Solar Energy*, vol. 195, pp. 685-696, 2020.

[48] I. A. Ibrahim and T. Khatib, "A novel hybrid model for hourly global solar radiation prediction using random forests technique and firefly algorithm," *Energy Conversion and Management*, vol. 138, pp. 413-425, 2017.

[49] R. H. Inman, H. T. Pedro and C. F. Coimbra, "Solar forecasting methods for renewable energy integration," *Progress in Energy and Combustion Science*, vol. 39, no. 6, December 2013, pp. 535-576, 2013.

[50] R.H. Inman, H.T. Pedro and C.F. Coimbra, "Solar forecasting methods for renewable energy integration," *Progress in Energy Combustion Science*, vol. 39, no. 6, pp. 535â€"576, 2013.

[51] S. Jacobsson and V. Lauber, "The politics and policy of energy system transformation – explaining the German diffusion of renewable energy technology," *Energy Policy*, vol. 34, no. 3, pp. 256-276, 2006.

[52] T. L. James, E. C. Brown and K. B. Keeling, "A hybrid grouping genetic algorithm for the cell formation problem," *Computers & Operations Research*, vol. 34, pp. 2059-2079, 2007.

[53] W. Ji and K. C. Chee, "Prediction of hourly solar radiation using a novel hybrid model of ARMA and TDNN," *Solar Energy*, vol. 85, no. 5, pp. 808-817, 2011.

[54] S. Jiménez-Fernández, C. Camacho-Gómez, R. Mallol-Poyato, J. C. Fernández, J. Del Ser, A. Portilla-Figueras and S. Salcedo-Sanz, "Optimal microgrid topology design and siting of distributed generation sources using a multi-objective substrate layer Coral Reefs Optimization algorithm," *Sustainability*, vol. 11, no. 1, pp. 1-21, 2019.

[55] G. John, R. Kohavi, and K. Pfleger, "Irrelevant features and the subset selection problem," in *Proceedings of the 11th International Conference on Machine Learning*, pp. 121-129, Morgan Kaufmann, 1994.

[56] S. A. Kalogirou, "Designing and Modeling Solar Energy Systems," *Solar Energy Engineering* (Second Edition), Chapter 11, pp. 583-699, 2014.

[57] T. Khatib, A. Mohamed amd K. Sopian, "A review of solar energy modeling techniques," *Renewable and Sustainable Energy Reviews*, vol. 16, pp. 2864-2869, 2012.

[58] T. Khatib, A. Mohamed and K. Sopian, “A review of solar energy modeling techniques,” *Renewable and Sustainable Energy Reviews*, vol. 16, pp. 2864-2869, 2012.

[59] S. Kirkpatrick and M. Vecchi, “Optimization by Simulated Annealing,” *Science*, vol. 220, pp. 671-680, 1983.

[60] J. Kleissl, “Solar Energy Forecasting and Resource Assessment,” *Academic Press*, 2013.

[61] V. B. Kreng and T. Lee, “Modular product design with grouping genetic algorithm – a case study,” *Computers & Industrial Engineering*, vol. 46, no. 3, pp. 443-460, 2004.

[62] R. Kohavi, and G. H. John, “Wrappers for features subset selection,” *International Journal of Digital Libraries*,vol. 1, pp. 108-121, 1997.

[63] H. Lan, C. Zhang, Y. Y. Hong, Y. He and S. Wen, “Day-ahead spatiotemporal solar irradiation forecasting using frequency-based hybrid principal component analysis and neural network,” *Applied Energy*, vol. 247, pp. 389-402, 2019.

[64] M. A. F. Lima, P. C. Carvalho, L. M. Fernández-Ramírez and A. P. Braga, “Improving solar forecasting using Deep Learning and Portfolio Theory integration,” *Energy*, vol. 195, no. 117016, 2020.

[65] M. Li, C. Miao and C. Leung, “A Coral Reef Algorithm Based on Learning Automata for the Coverage Control Problem of Heterogeneous Directional Sensor Networks,” *Sensors*, vol. 15, pp. 30617-30635, 2015.

[66] A. Linares-Rodriguez, J. A. Ruiz-Arias, D. Pozo-Vazquez and J. Tovar-Pescador, “An artificial neural network ensemble model for estimating global solar radiation from Meteosat satellite images,” *Energy*, vol. 61, pp. 636-645, 2013.

[67] G. López, B. Batlles and J. Tovar-Pescador, “Selection of input parameters to model direct solar irradiance by using artificial neural networks,” *Energy*, vol. 30, pp. 1675-1684, 2005.

[68] N. Lu, J. Qin, K. Yang and J. Sun, “A simple and efficient algorithm to estimate daily global solar radiation from geostationary satellite data,” *Energy*, vol. 36, pp. 3179-3188, 2011.

[69] A. Mellit and S. A. Kalogirou, “Artificial intelligence techniques for photovoltaic applications: a review,” *Progress in Energy and Combustion Science*, vol. 34, no. 5, pp. 574-632, 2008.

[70] K. Mohammadi, S. Shamshirband, C. W. Tong, M. Arif, D. Petkovic and S. Che, “A new hybrid support vector machine-wavelet transform approach for estimation of horizontal global solar radiation,” *Energy Conversion and Management*, vol. 92, pp. 162-171, 2015.

[71] K. Mohammadi, S. Shamshirband, D. Petkovic and H. Khorasanizadeh, “Determining the most important variables for diffuse solar radiation prediction using adaptive neuro-fuzzy methodology; case study: City of Kerman, Iran,” *Renewable and Sustainable Energy Reviews*, vol. 53, pp. 1570-1579, 2016.

[72] A. Mosavi, M. Salimi, S. F. Ardabili, T. Rabczuk, S. Shamshirband annd A. R. Varkonyi-Koczy, “State of the art of machine learning models in energy systems, a systematic review,” *Energies*, vol. 12, no. 1301, 2019.

[73] S. M. Mousavi, E. S. Mostafavi and P. Jiao, “Next generation prediction model for daily solar radiation on horizontal surface using a hybrid neural network and simulated annealing method,” *Energy Conversion and Management*, vol. 153, pp. 671-682, 2017.

[74] J. Mubiru, “Predicting total solar irradiation values using artificial neural networks,” *Renewable Energy*, vol. 33, pp. 2329-2332, 2008.

[75] L. Olatomiwa, S. Mekhilef, S. Shamshirband, K. Mohammadi, D. Petkovic and S. Ch, “A support vector machine-firefly algorithm-based model for global solar radiation prediction,” *Solar Energy*, vol. 115, pp. 632-644, 2015.

[76] P. A. Owusu and S. Asumadusarkodie, “A review of renewable energy sources, sustainability issues and climate change mitigation,” *Cogent Engineering*, vol. 3, no. 1167990, 2016.

[77] Z. Pang, F. Niu and Z. O’Neill, “Solar radiation prediction using recurrent neural network and artificial neural network: A case study with comparisons,” *Renewable Energy*, vol. 156, pp. 279-289, 2020.

[78] C. Paoli, C. Voyant, M. Muselli and M. L. Nivet, “Forecasting of preprocessed daily solar radiation time series using neural networks,” *Solar Energy*, vol. 84, pp. 2146-2160, 2010.

[79] J. Perdigao, R. Salgado, C. Magarreiro, P. M. Soares, M. J. Costa and H. P. Dasari, “An Iberian climatology of solar radiation obtained from WRF regional climate simulations for 1950 to 2010 period,” *Atmospheric Research*, vol. 198, pp. 151-162, 2017.

[80] R. Perez, E. Lorenz, S. Pelland, M. Beauharnois, G. Van Knowe, K. Hemker, et al. “Comparison of numerical weather prediction solar irradiance forecasts in the US, Canada and Europe,” *Solar Energy*, vol. 94, pp. 305-326, 2013.

[81] X. Qing and Y. Niu, “Hourly day-ahead solar irradiance prediction using weather forecasts by LSTM,” *Energy*, vol. 148, pp. 461-468, 2018.

[82] A. Rahimikoob, “Estimating global solar radiation using artificial neural network and air temperature data in a semi-arid environment,” *Renewable Energy*, vol. 35, pp. 2131-2135, 2010.

[83] M. Rana, I. Koprinska and V. G. Agelidis, “Univariate and multivariate methods for very short-term solar photovoltaic power forecasting,” *Energy Conversion and Management*, vol. 121, pp. 380-390, 2016.

[84] S. Rehman and M. Mohandes, “Artificial neural network estimation of global solar radiation using air temperature and relative humidity,” *Energy Policy*, vol. 36, no. 2, pp. 571-576, 2008.

[85] https://www.iea.org/reports/renewable-energy-market-update

[86] G. Sadeghi, A. L. Pisello, H. Safarzadeh, M. Poorhossein and M. Jowzi, “On the effect of storage tank type on the performance of evacuated tube solar collectors: Solar radiation prediction analysis and case study,” *Energy*, vol. 198, no. 117331, 2020.

[87] M. Sahin, Y. Kaya, M. Uyar and S. Yidirim, “Application of extreme learning machine for estimating solar radiation from satellite data,” *International Journal of Energy Research*, vol. 38, no. 2, pp. 205-212, 2014.

[88] S. Salcedo-Sanz, M. Prado-Cumplido, F. Pérez-Cruz, and C. Bousoño-Calzón, “Feature selection via genetic optimization,” in *International Conference on Artificial Neural Networks, ICANN2002*, Madrid, Spain., pp. 547–552, *Lecture Notes in Computer Science*, Springer-Verlag, 2002.

[89] S. Salcedo-Sanz, A. M Pérez-Bellido, E. G. Ortiz-García, A. Portilla-Figueras and L. Prieto, “Hybridizing the fifth generation mesoscale model with artificial neural networks for short-term wind speed prediction,” *Renewable Energy*, vol. 34, pp. 1451-1457, 2009.

[90] S. Salcedo-Sanz, A. M Pérez-Bellido, E. G. Ortiz-García, A. Portilla-Figueras, L. Prieto and F. Correoso, “Accurate short-term wind speed prediction by exploiting diversity in input data using banks of artificial neural networks,” *Neurocomputing*, vol. 72, no. 4, pp. 1336-1341, 2009.

[91] S. Salcedo-Sanz, C. Casanova-Mateo, A. Pastor-Sánchez, D. Gallo-Marazuela, A. Labajo-Salazar and A. Portilla-Figueras, “Direct solar radiation prediction based on Soft-Computing algorithms including novel predictive atmospheric variables,” *Intelligent Data Engineering and Automated Learning - IDEAL 2013*, Lecture Notes in Computer Science, vol. 8206, pp. 318-325, 2013.

[92] S. Salcedo-Sanz, C. Casanova-Mateo, J. Muñoz-Marí and G. Camps-Valls, “Efficient prediction of daily global solar irradiation using temporal Gaussian Processes,” *IEEE Geoscience and Remote Sensing Letters*, vol. 11, pp. 1136-1140, 2014.

[93] S. Salcedo-Sanz, A. Pastor-Sánchez, L. Prieto, A. Blanco-Aguilera and R. García-Herrera, “Feature selection in wind speed prediction systems based on a hybrid coral reefs optimization–Extreme learning machine approach,” *Energy Conversion and Management*, vol. 87, pp. 10-18, 2014.

[94] S. Salcedo-Sanz, C. Casanova-Mateo, A. Pastor-Sánchez and M. Sánchez-Girón, "Daily Global Solar Radiation Prediction based on a Hybrid Coral Reefs Optimization – Extreme Learning Machine Approach," *Solar Energy*, vol. 105, pp. 91-98, 2014.

[95] S. Salcedo-Sanz C. Camacho-Gómez, A. Magdaleno, E. Pereira and A. Lorenzana, "Structures vibration control via Tuned Mass Dampers using a co-evolution Coral Reefs Optimization algorithm," *Journal of Sound and Vibration*, vol. 393, pp. 62-75, 2017.

[96] S. Salcedo-Sanz, L. Cornejo-Bueno, L. Prieto, D. Paredes and R. García-Herrera, "Feature selection in machine learning prediction systems for renewable energy applications," *Renewable and Sustainable Energy Reviews*, vol. 90, pp. 728-741, 2018.

[97] S. Salcedo-Sanz, R. C. Deo, L. Cornejo-Bueno, C. Camacho-Gómez and S. Ghimire, "An efficient neuro-evolutionary hybrid modelling mechanism for the estimation of daily global solar radiation in the Sunshine State of Australia," *Applied Energy*, vol. 209, pp. 79-94, 2018.

[98] S. Salcedo-Sanz, P. Ghamisi, M. Piles, M. Werner, L. Cuadra, A. Moreno-Martínez, E. Izquierdo-Verdiguier, J. Muñoz-Marí, A. Mosavi and G. Camps-Valls, "Machine learning information fusion in Earth observation: A comprehensive review of methods, applications and data sources," *Information Fusion*, vol. 63, pp. 256-272, 2020.

[99] S. Salcedo-Sanz, J. Del Ser, I. Landa-Torres, S. Gil-López and J.A. Portilla-Figueras, "The Coral Reefs Optimization algorithm: a novel metaheuristic for efficiently solving optimization problems," *Science World Journal*, vol. 739768, 2014.

[100] S. Salcedo-Sanz, D. Gallo-Marazuela, A. Pastor-Sánchez, L. Carro-Calvo, J. A. Portilla-Figueras and L. Prieto, "Offshore wind farm design with the Coral Reefs Optimization algorithm," *Renewable Energy*, vol. 63, pp. 109-115, 2014.

[101] S. Salcedo-Sanz, C. Camacho-Gómez, D. Molina and F. Herrera, "A coral reefs optimization algorithm with substrate layers and local search for large scale global optimization," *IEEE Congress on Evolutionary Computation (CEC)*, pp. 3574-3581, 2016.

[102] S. Salcedo-Sanz, A. Pastor-Sánchez, D. Gallo-Marazuela, and A. Portilla-Figueras, "A Novel Coral Reefs Optimization Algorithm for Multi-objective Problems," *Intelligent Data Engineering and Automated Learning Conference*, LNCS, vol. 8206, pp. 326-333, 2013.

[103] S. Salcedo-Sanz, A. Pastor-Sánchez, A. Portilla-Figueras and L. Prieto, "Effective multi-objective optimization with the coral reefs optimization algorithm," *Engineering Optimization*, vol. 48, no. 6, 2016.

[104] S. Salcedo-Sanz, J. Muñoz-Bulnes and M. Vermeij, "New Coral Reefs-based Approaches for the Model Type Selection Problem: A Novel Method to Predict a Nation's Future Energy Demand," *International Journal of Bio-inspired Computation*, vol. 10, no.3, pp. 145-158, 2017.

[105] S. Salcedo-Sanz, “A review on the Coral Reefs Optimization algorithm: new development lines and current applications,” *Progress in Artificial Intelligence*, vol. 6, no. 1, pp 1-15, 2017.

[106] S. Salcedo-Sanz, C. Camacho-Gómez, R. Mallol-Poyato, S. Jiménez-Fernández and J. Del Ser, “A novel Coral Reefs Optimization algorithm with substrate layers for optimal battery scheduling optimization in micro-grids,” *Soft Computing*, vol. 20, no. 11, pp. 4287-4300, 2016.

[107] S. Salcedo-Sanz, R. García-Herrera, C. Camacho-Gómez, E. Alexandre, L. Carro-Calvo and F. Jaume-Santero, “Near-optimal selection of representative measuring points for robust temperature field reconstruction with the CRO-SL and analogue methods,” *Global and planetary change*, vol. 178, pp. 15-34, 2019.

[108] R. Sánchez-Montero, C. Camacho-Gómez, P. L. López-Espí and S. Salcedo-Sanz, “Optimal Design of a Planar Textile Antenna for Industrial Scientific Medical (ISM) 2.4 GHz Wireless Body Area Networks (WBAN) with the CRO-SL Algorithm,” *Sensors Journal*, vol. 18, pp. 1-17, 2018.

[109] Z. Sen, Solar Energy Fundamentals and Modeling Techniques, Atmosphere, Environment, Climate change and renewable energy, London, Springer-Verlag, 2008.

[110] O. Senkal and T. Kuleli, “Estimation of solar radiation over Turkey using artificial neural network and satellite data,” *Applied Energy*, vol. 86, no. 7-8, pp. 1222-1228, 2009.

[111] W. C Skamarock, J. B. Klemp, J. Dudhia, D. O. Gill, D. M. Barker, W. Wang and J. G. Powers, A description of the advanced research WRF version 2, Technical report, National Center For Atmospheric Research Boulder Co. Mesoscale and Microscale Meteorology Div, 2005.

[112] K. H. Solangi, M. R. Islam, R. Saidur R, N. A. Rahim and H. Fayaz, “A review on global solar energy policy,” *Renewable and Sustainable Energy Reviews*, vol. 15, pp. 2149â€“2163, 2011.

[113] S. Solorio-Fernández, J. A. Carrasco-Ochoa and J. F. Martínez-Trinidad, “A new hybrid filter-wrapper feature selection method for clustering based on ranking,” *Neurocomputing*, vol. 214, pp. 866-880, 2016.

[114] A. Sozen, E. Arcakliogblu and M. Ozalp, “Estimation of solar potential in Turkey by artificial neural networks using meteorological and geographical data,” *Energy Conversion and Management*, vol. 45, pp. 3033-3052, 2004.

[115] K. Torkkola and William M. Campbell, “Mutual information in learning feature transformations,” in *Proc. 17th International Conf. on Machine Learning*, pp. 1015-1022, Morgan Kaufmann, San Francisco, CA, 2000.

[116] K. Torkkola, “On feature extraction by mutual information maximization,” in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, pp. 821-824, 2002.

[117] C. W. Tsai, W. Y. Chang, Y. C. Wang and H. Chen, “A high-performance parallel coral reef optimization for data clustering,” *Soft Computing*, vol. 23, pp. 9327-9340, 2019.

[118] R. Urraca, T. Huld, A. Gracia-Amillo, F. J. Martinez-de-Pison, F. Kaspar and A. Sanz-Garcia, “Evaluation of global horizontal irradiance estimates from ERA5 and COSMO-REA6 reanalyses using ground and satellite-based data”, *Solar Energy*, vol. 164, pp. 339-354, 2018.

[119] C. Voyant, M. Musellia, C. Paolia, M.L. Niveta, “Optimization of an artificial neural network dedicated to the multivariate forecasting of daily global radiation,” *Energy*, vol. 36, no. 1, pp. 348-359, 2011

[120] C. Voyant, M. Muselli, C. Paoli and M. L. Nivet, “Hybrid methodology for hourly global radiation forecasting in Mediterranean area,” *Renewable Energy*, vol. 53, pp. 1-11, 2013.

[121] C. Voyant, G. Notton, S. Kalogirou, M. L. Nivet, C. Paoli, F. Motte and A. Fouilloy, “Machine learning methods for solar radiation forecasting: A review,” *Renewable Energy*, vol. 105, pp. 569-582, 2017.

[122] R. Yacef, M. Benghanem and A. Mellit, “Prediction of daily global solar irradiation data using Bayesian neural network: A comparative study,” *Renewable Energy*, vol. 48, pp. 146-154, 2012.

[123] A. K. Yadav, and S. S. Chandel, “Solar radiation prediction using Artificial Neural Network techniques: A review,” *Renewable and Sustainable Energy Reviews*, vol. 33, pp. 772-781, 2014.

[124] A. K. Yadav, H. Malik and S.S. Chandel, “Selection of most relevant input parameters using WEKA for artificial neural network based solar radiation prediction models,” *Renewable and Sustainable Energy Reviews*, vol. 31, pp. 509-519, 2014.

[125] C. Yan, J. Ma, H. Luo and A. Patel, “Hybrid binary Coral Reefs Optimization algorithm with Simulated Annealing for Feature Selection in high-dimensional biomedical datasets,” *Chemometrics and Intelligent Laboratory Systems*, vol. 184, pp. 102-111, 2019.

[126] J. Yang and V. Honavar, “Feature subset selection using a genetic algorithm,” *IEEE Intelligent Systems*, vol. 13, no. 2, pp. 44-49, 1998.

[127] X. Yao, Y. Liu and G. Lin, “Evolutionary programming made faster,” *IEEE Transactions on Evolutionary computation*, vol. 3, no. 2, pp. 82-102, 1999.

[128] J. M. Yeom, R. C. Deo, J. F. Adamwoski, T. Chae, D. S. Kim, K. S. Han and D. Y. Kim, "Exploring solar and wind energy resources in North Korea with COMS MI geostationary satellite data coupled with numerical weather prediction reanalysis variables," *Renewable and Sustainable Energy Reviews*, vol. 119, no. 109570, 2020.

[129] H.J.J. Yu, N. Popiolek and P. Geoffron, "Solar photovoltaic energy policy and lobalization: a multiperspective approach with case studies of Germany, Japan, and China," *Progress in Photovoltaics: Research and Applications*, vol. 24, no. 4, pp. 458-476, 2016.

[130] F. Wang, Z. Zhen, Z. Mi, H. Sun, S. Su and G. Yang, "Solar irradiance feature extraction and support vector machines based weather status pattern recognition model for short-term photovoltaic power forecasting," *Energy and Buildings*, vol. 86, pp. 427-438, 2015.

[131] H. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio and V. Vapnik,"Feature selection for SVMs," *Advances in NIPS 12*, MIT Press, pp. 526-532, 2000.

[132] A. Will, J. Bustos, M. Bocco,J. Gotaya and C. Lamelas, "On the use of niching genetic algorithms for variable selection in solar radiation estimation," *Renewable Energy*, vol. 50, pp. 168-176, 2011.

[133] J. Wu and C. K. Chan, "Prediction of hourly solar radiation using a novel hybrid model of ARMA and TDNN," *Solar Energy*, vol. 85, pp. 808-817, 2011.

[134] H. Zang, L. Cheng, T. Ding, K. W. Cheung, M. Wang, Z. Wei and G. Sun, "Application of functional deep belief network for estimating daily global solar radiation: A case study in China," *Energy*, vol. 191, no. 116502, 2020.

[135] L. F. Zarzalejo, L. Ramirez and J. Polo, "Artificial intelligence techniques applied to hourly global irradiance estimation from satellite-derived cloud index," *Energy*, vol. 30, pp. 1685-1697, 2005.

[136] J. Zeng and W. Qiao, "Short-term solar power prediction using a support vector machine," *Renewable Energy*, vol. 52, pp. 118-127, 2013.

[137] S. Zhang and Y. He, "Analysis on the development and policy of solar PV power in China," *Renewable and Sustainable Energy Reviews*, vol. 21, pp. 393-401, 2013.

[138] R. Zhang, Y. Lan, G. B. Huang and Z. Xu, "Universal Approximation of Extreme Learning Machine With Adaptive Growth of Hidden Nodes," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, pp. 365-371, 2012.

Part VI

Copy of the Publications

Applied Soft Computing 43 (2016) 619–632

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

A Lamarckian Hybrid Grouping Genetic Algorithm with repair heuristics for resource assignment in WCDMA networks☆

L. Cuadra[a], A. Aybar-Ruíz[a], M.A. del Arco[a], J. Navío-Marco[b], J.A. Portilla-Figueras[a], S. Salcedo-Sanz[a,*]

[a] Department of Signal Processing and Communications, Universidad de Alcalá, Alcalá de Henares, Spain
[b] Department of Business Organization, UNED, Spain

ARTICLE INFO

Article history:
Received 8 July 2015
Received in revised form 6 December 2015
Accepted 27 January 2016
Available online 18 February 2016

Keywords:
Lamarckian Hybrid Grouping Genetic Algorithm
WCDMA networks
Resource assignment

ABSTRACT

In this paper we propose a novel Lamarckian Hybrid Grouping Genetic Algorithm (LHGGA) with repair heuristics for a problem of resources assignment to mobile terminals (or, simply, users) in Wide-band Code Division Multiple Access (WCDMA) networks. We propose a novel problem formulation that takes into account all the interference terms, which strongly depend on the assignment to be done. The second contribution is a cost function (to be minimized) with weighted components, which is composed of not only the load factors (including the mentioned interference terms) but also other utilization ratios for aggregate capacity, codes, power, and users without service. The second group of contributions is related to the LHGGA approach. On the one hand, we propose a novel encoding scheme, suitable for the novel problem formulation. On the other hand, we present fully tailored operators. We emphasize the proposal of a repair operator of unphysical candidates (which are substituted by their repaired versions), and a crossover operator, able to acts on groups (users assigned to a base station) in a very efficient way. The proposed LHGGA exhibits a superior performance than that of the conventional method, since most of users receive the demanded services along with a more efficient use of resources per user. The LHGGA approach has been successfully applied to a variety of scenarios: different number of users, distributions, or users profiles.

1. Introduction

According to the Global mobile Suppliers Association (GSA) there are currently 6.44 billion subscriptions worldwide using 3rd Generation Partnership Project (3GPP) mobile networks [1]: Global System for Mobile Communications (GSM) networks, Wide-band Code Division Multiple Access (WCDMA) networks (also known Third Generation (3G) networks), and Forth Generation (4G) cellular networks [1–3], such as Long Term Evolution (LTE) [4,5]. To provide the best customer experience, these networks form a mobile access "ecosystem of Heterogeneous Networks". The term Heterogeneous Network (HetNet) [6] is often used to name a global network that consists of several mobile network technologies (GSM, WCDMA and LTE) covering the same geographical area [7], and also to describe the coexistence of different non-homogenous cells (macrocells, microcells, femtocells [8–10]) to assure high data rates in small areas with large densities of users.

High Speed Packet Access (HSPA), based on WCDMA, is the most widely used deployed mobile *broadband* technology in the world. In fact, HSPA is not a unique technology, but a set of technologies that allow mobile operators to easily *upgrade* their already deployed WCDMA networks to support a very efficient provision of speech services and mobile broadband data services (high speed Internet access, music-on-demand, and TV and video streaming, to name just a few). Currently, 83% of mobile operators worldwide are investing and upgrading their WCDMA-based 3G networks [11] and have 1.83 billion WCDMA subscribers [1]. WCDMA/HSPA technology is expected to cover 90% of the world's population by 2020, serving about 3.8 billion subscribers [12]. These figures illustrate the importance of properly planning and dimensioning WCDMA networks. The question that motivates this work is how to assign the limited WCDMA resources to mobile terminals (users equipments, or, simply, users).

One of these telecommunication resources are the available frequencies. In WCDMA cellular networks, a number of users are allowed to utilize simultaneously the same frequency. To separate the communications, the network assigns a "channelization code" to each communication, so that only the corresponding receiver is able to extract the information that has been sent to it. However, a given amount of interference appears between communication links using the same frequency. A parameter called "load factor" is commonly used to quantify the influence of interference. It is defined as the ratio between the interference and the total perturbation (thermal noise + interference) [13–16]. The most used conventional approach for dimensioning WCDMA networks is based on keeping the interference and load

☆ This paper is an improved version of the paper "A Novel Grouping Genetic Algorithm for Assigning Resources to Users in WCDMA Networks", presented at EvoComNet2015 and published in Applications of Evolutionary Computation: 18th European Conference, EvoApplications 2015, Copenhagen, Denmark, April 8–10, 2015, Proceedings (Vol. 9028, p. 42). Springer.

* Corresponding author at: Department of Signal Processing and Communications, Universidad de Alcalá, 28871 Alcalá de Henares, Madrid, Spain.
Tel.: +34 91 885 6731; fax: +34 91 885 6699.
E-mail address: sancho.salcedo@uah.es (S. Salcedo-Sanz).

http://dx.doi.org/10.1016/j.asoc.2016.01.046
1568-4946/© 2016 Elsevier B.V. All rights reserved.

factor lower than a given characteristic thresholds [14,17]. The problem of assigning users to the serving base station (BS) – also known "cell selection" [18] – is classically tackled using algorithms that are based on the selection of the cell with the minimum propagation loss and/or leading to the maximal signal-to-interference plus noise ratio (SINR).

However, the current ever increasing demand of higher speeds in mobile communications [12] reveals that there are *other resources* that should be taken into account. One of the most evident is based on the fact that the aggregation of increasing numbers of users with higher data rates is leading to a bottleneck in the aggregation interface at the BS, in the sense that these aggregated rates could be higher than the available backhaul capacity [19,20]. Other limiting resources are the number of channelization codes (whose number is limited by the way they are generated [13]) and the available power of base stations. In this respect, we have recently tackled the problem of assigning these WCDMA resources by using a modified version of a Grouping Genetic Algorithm (GGA) [21], which made used of *approximated* expressions of the load factors (since these did *not* take into account all the interferences). Notwithstanding, this approximation, which models the inferences arriving from other cells as an average value, is useful and is often considered when dimensioning WCDMA networks [13–17].

In the present work we *do* take into account *all* the interferences, and proceed further by proposing a Hybrid Grouping Genetic Algorithm (HGGA) with *repair heuristics* [22] to manage the creation of unfeasible individuals. In general, the term "hybrid" is applied to Evolutionary Algorithms (EA) –and to GGA, in particular– when the repair method is used as a constraint handling procedure to reduce the search space by only considering those feasible individuals. Note that, according to [22], the purpose of a hybrid algorithm is different from that of a memetic algorithm [23]: whereas the local search in memetic algorithms is focused on the improvement of the fitness of an individual, the repair method in a hybrid algorithm aims to *not* violate constraints, by reducing the search space only to feasible individuals. Hybrid approaches with repair heuristics can be classified into "Lamarckian" and "Baldwinian" approaches [24]. If the unfeasible chromosome is substituted by its repaired version after the application of the local repair heuristic, the algorithm is called Lamarckian [25]. Baldwinian [26] hybrid algorithms are those in which the individual population does not change after the application of the repair heuristic, only the fitness function being modified [22]. As will be shown later on, our approach is a Lamarckian HGGA.

With this in mind, the purpose of this work is to explore the feasibility of a Lamarckian Hybrid Grouping Genetic Algorithm (LHGGA) with repair heuristics [22] to near-optimally assign the WCDMA resources (aggregated capacity, power, codes) of M base stations to N users, by minimizing a cost function composed of *weighted* constituents such as the load factors (which include, as a novelty, a detailed modeling of *all* possible interference signals), the fractions of available resources (aggregated capacity, power, codes), and the fraction of users *without service*. This is important because a reduced number of users without service is perceived by customers as high service availability, which help operators to increase market share.

The present research work differs from our previous work [21] in:

(1) We model and compute the load factors considering *all* the interference, which, as will be shown, are different depending on the base station any user is assigned to.
(2) We propose an LHGGA with repair heuristics, which is an improved version of the GGA explored in [21], since the LHGGA implementation is able to repair chromosomes encoding unphysical.
(3) The cost function to be minimized is more flexible than that of [21] in the sense that its constituent elements (load factors, fractions of available resources – aggregated capacity, power, codes – and fraction of users without service) may be multiplied by weight factors. This helps the designer prioritize one or several constituents. For instance, aiming at reducing the fraction of users without service, a higher weight factor can be assigned to the fraction of users without service.
(4) We have carried out a completely novel and more extensive set of experiments. On the one hand, we compare the LHGGA performance to that of a conventional approach (CA) that only minimizes the load factors. On the other hand, we have used the LHGGA approach to study the assignments in different scenarios: increasing number of users, different user distributions, or changes in users profiles, with non-uniform traffic patterns. In all cases, the algorithm predicts situations empirically proven by the experience of the operators.

The structure of the rest of this paper is as follows. While Section 2 reviews the related works, Section 3 summarizes some WCDMA fundamental aiming at better explaining the problem statement and different approaches. Section 4 focuses on describing in detail the our problem formulation (including detail models of the affecting interferences), along with a characterization of the resources to be assigned. Section 5, which is the soft-computing core of our work, describes the LHGGA we propose to tackle the aforementioned problem, emphasizing the repair heuristics. We also focus on detailing the LHGGA encoding and the different crossover and mutation operators implemented. Section 6 shows the experimental work we have carried out in order to show the good performance of the proposed LHGGA. Finally, Section 7 completes the paper by discussing the main findings obtained in this work.

2. Related work

We have mentioned that the problem of assigning users to base stations (or cell selection problem [18]) can be tackled using classical algorithms based on the selection of the cell with the minimum propagation loss and/or leading to the maximal signal-to-interference plus noise ratio (SINR). One of these cell selection algorithms is the "Best-Server Cell Selection" (BSCS) algorithm [3]. In this strategy, users are always assigned to the BS with the lowest propagation loss. This BS is usually called "best base station" (BBS) or best server (BSV). Although this algorithm leads to an efficient use of radio resources, however it suffers from inefficiencies because the aggregate capacity of the BBS could be saturated ("overloaded"). Other user assignment algorithm used in WCDMA networks is the "Radio Prioritized Cell Selection" (RPCS) algorithm [3]. In this algorithm, a list of candidate BSs is made as follows: all the BSs having a difference in propagation loss (with respect to the best base station) lower than a given propagation loss margin (PLM) are considered as candidate BSs. Then, among the candidate BSs whose capacity is not overloaded, the conventional RPCS algorithm selects the BS having the minimum propagation loss, and the user is assigned to it. This algorithm takes into account capacity limits, but it comes at the expense of radio degradation because of the potential selection of non-optimal cells.

A very interesting approach to optimize both the radio interface and the backhaul capacity have been recently explored in [19] using a cell selection algorithm called "Transport Prioritized Cell Selection" (TPCS) for any user u_j. It works like the RPCS algorithm when the already used capacity of all the BSs in the list of candidate BSs is lower than a certain threshold. However, when the used capacity of at least one of the candidate BSs is higher than such threshold, the TPCS algorithm prioritizes BSs according to their capacity occupancy. The authors made use of an analytical model based on multi-dimensional Markov chains to assess the performance of the TPCS algorithm, and validated its results using a Monte Carlo algorithm. The results pointed out that this approach was useful to achieve a more efficient use of backhaul capacity (when compared to classical BSCS and RPCS cell selection algorithms). In a similar line of research, [27] focused on the problem of base station assignment in Orthogonal Frequency-Division Multiple Access (OFDMA) cellular networks, and proposed a heuristic that made use of Lagrange multipliers, leading to the conclusion that the algorithm was able to give the same capacity but using less backhaul resources.

For comparative purposes, Table 2 lists the pros. and cons. of these methods when compared to the one we propose in this paper and in our previous, simplified approach [21]. In that work, we proposed a GGA [28–30] to assign resources (aggregate capacity, power, codes) to users in WCDMA networks, assuming simplified versions of the load factors (the usual in text such as [13–16]), and we did *not* make use of repair heuristics. The present work differs from [21] in the contributions listed (1)–(4) in Section 1. Although in a different approach from that in [21], the GGA concept has been already applied to other telecommunication problems such as mobile communication network design [31–33], or OFDMA-based multicast wireless systems [34].

Besides the proposal to optimize the radio interface and the backhaul capacity [19], there are also some works, which are only *partially* related to the underlying problem (focused only on the jointly assignment of users to base stations and power [35,36], the base stations and beam-forming schemes [37,38], or automatic procedures for the design of WCDMA networks [39]). However, apart from our preliminary approach [21], there appears to be *no* study that combines all the factors (load factors and interferences, backhaul capacity, power constrains, or number of codes) using Soft Computing (SC) approaches.

Table 1
List of acronyms used in this paper.

Acronym service	Meaning
3G	Third Generation
3GPP	Generation Partnership Project
4G	Forth Generation
AIS	Artificial Immune System
BBS	best base station
BS	Base Station
BSCS	Best-Server Cell Selection
BSV	Best server
EA	Evolutionary Algorithms
GA	Genetic Algorithm
GEE	Grouping Evolutionary Strategy
GGA	Grouping Genetic Algorithm
GHS	Grouping Harmony Search
GP	Genetic Programing
GPS	Grouping Particle Swarm Optimization algorithm
GSA	Global mobile Suppliers Association
GSM	Global System for Mobile Communications
HetNet	Heterogeneous Network
HGGA	Hybrid Grouping Genetic Algorithm
HSDPA	High-Speed Downlink Packet Access
HSPA	High Speed Packet Access
LHGGA	Lamarckian Hybrid Grouping Genetic Algorithm
LTE	Long Term Evolution
nB	Node B
OFDMA	Orthogonal Frequency-Division Multiple Access
OVSF	Orthogonal Variable Spreading Factor
PLM	Propagation Loss Margin
PM	Propagation model
PSO	Particle Swarm Optimization
RPCS	Radio Prioritized Cell Selection
SC	Soft Computing
SI	Swarm Intelligence
SP	Service Profile
SNIR	Signal to noise-and-interference ratio
TPCS	Transport Prioritized Cell Selection
WCDMA	Wide-band Code Division Multiple Access

SC approaches have dealt with other problems related to 3G mobile networks optimization, although with different purposes, such as [40], in which an evolutionary-based approach has been proposed to cell size determination in the context of WCDMA networks, taking into account different number of users and services provided by the system. In other approach, Particle Swarm Optimization (PSO) has also been explored to tackle the problem of base station configuration for planning WCDMA networks [41]. Genetic algorithms (GA) have been used widely in WCDMA networks [42–45], for instance, for the problem of codes allocation [43], and to optimize the location and configuration of base stations in WCDMA network planning [44]. Genetic algorithms have also been applied to the deployment of base stations, taking into account capacity and coverage in WCDMA networks and using different antenna heights [45]. The soft-computing approach of Genetic Programing (GP) has been explored as a promising method for automated optimization design of base stations in WCDMA networks [46]. A hybrid optimization systems based on Swarm Intelligence (SI) has been applied to multi-user scheduling in HSDPA (High-Speed Downlink Packet Access) within 3G networks [47]. Artificial Immune System (AIS) algorithms have also been applied to 3G network optimization problems, just like in [48], where an artificial immune system has been used to solve a twofold problem in which the users admission and control are considered. Recently, [49] has explored an evolutionary multi-objective algorithm for WCDMA network planning which includes an iterative power control method with the simplification of neglecting the interference arising from channels without no-load coverage.

As shown, although there is a considerable variety of research works related to a greater or lesser degree to our proposal, to the best of our knowledge, there appears to be *no* study that: (1) formulates the assignment problem involving load factors and interferences, backhaul capacity, power constrains, number of codes, and number of users; and (2) tackles it by using an LHGGA with repair heuristic aiming at finding near-optimal solutions to the problem at hand.

3. WCDMA background

We have mentioned in Section 1 that users in WCDMA networks are allowed to use simultaneously the same electromagnetic carrier f_C. To separate two communications on the same carrier, the network assigns a channelization code to each communication. This is done by multiplying the user data (with bit rate R_b) by a code or sequence of special bits (called "chips"), whose rate ("chip rate" W) is a characteristic network parameter (W = 3.84 Mcps) much higher than that of the user bit rate $W \gg R_b$ [14]. This concept has been represented in Fig. 1. In this example, the code assigned to this communication is $\{1, -1, 1, -1, -1, 1, -1, 1\}$. Note in Fig. 1 that each *bit* of the user's signal ($d(t)$) is multiplied by the code sequence $\{1, -1, 1, -1, -1, 1, -1, 1\}$. In WCDMA networks, Orthogonal Variable Spreading Factor (OVSF) codes, which were originally proposed in [50], are used as channelization codes. These codes are generated from a code trees, based on the required data rates (R_b)[50–52].

Although orthogonal property helps ideally reduce interference, however, the remaining communications using the same frequency become somehow interference signals. This is illustrated in Fig. 2. The dashed area represents the *cell* that is covered by a BS or "node B" (nB), in WCDMA terminology. Throughout this work, both words will be used interchangeably. The nB labeled B_k in Fig. 2 will be used as a "reference" throughout this paper aiming at better explaining the most complex aspects of the involved interference terms. $n_u^{B_k}$ represents the number of user that the base station B_k is serving. In particular, a reference user, u_l, assigned to B_k, has also been represented. User u_l, which emits a power $p_e(l)$, will be used later on to explain how interference is calculated. $p_{\mathrm{R},B_k}(l)$ represents the power received at the base station B_k emitted by user u_l. The total interference must contains not only those interferences generated by the users in the own cell (for instance, user u_j in Fig. 2) but also those arising from other users located in other cells (user u_m). Note that apart from the interferences appearing in the uplink (UL) –signals moving from the users to the BS– there are also others in the downlink (DL). A representative example is the interference produced by the base station B_q ($q \neq k$), which interferes on the

Table 2
Comparison among different cell selection algorithms (or user assignment algorithms). See Table 1 for acronyms.

Method	Pros	Cons
BSCS	• Efficient use of radio resources	• Base station capacity can be overloaded [19] • Considers neither codes, nor power, nor users without service
RPCS	• Includes BS capacity limits [19]	• Radio degradation because of potential selection of non-optimal cells • Considers neither codes, nor power, nor users without service
TPCS	• Optimizes both radio interface and backhaul capacity [19]	• Considers neither codes, nor power, nor users without service
HLGGA	• Able to optimize radio resources, capacity, codes, power, and users	• More complex interactions

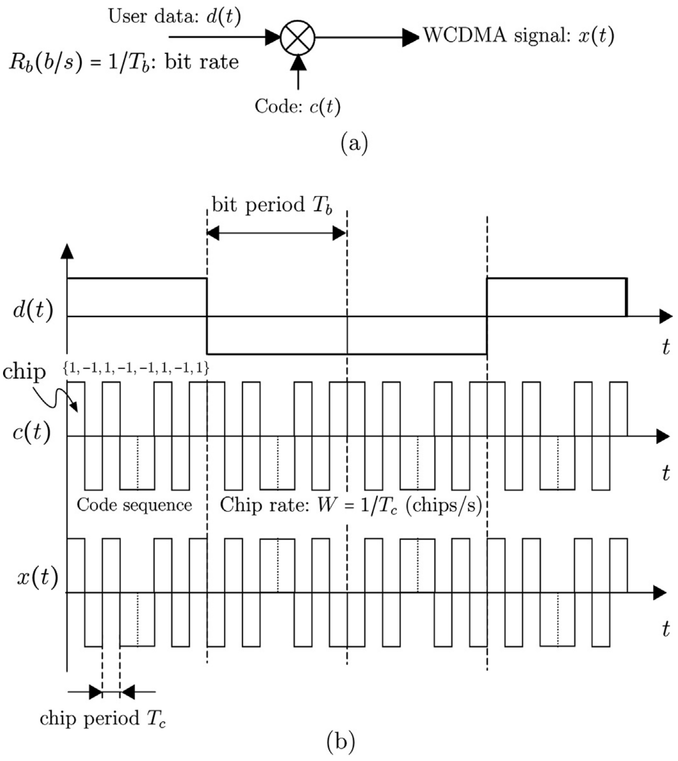


Fig. 1. (a) Simplified example of WCDMA signal generation. (b) Each bit of the user's data, $d(t)$, with a bit rate $R_b(b/s) = 1/T_b$ (T_b being the bit period), is multiplied by the code $c(t)$ assigned to such communication. In this example, the code sequence is {1,−1,1,−1,−1,1,−1,1}. T_c is the chip period and $W = 1/T_c$ is the chip rate.

DL signal corresponding to the link $u_l \leftrightarrow B_k$, involving the serving B_k and the reference user u_l.

The load factor in all the up-links of the cell served by the nB B_k – defined as the ratio between the interference and the *total* noise (thermal + interference) [3,53,54] – can be estimated as

$$\eta_{\mathrm{UL}}(B_k) \approx (1+\xi)\cdot \sum_{j=1}^{n_u^{B_k}} \frac{1}{1+\frac{1}{(e_b/n_0)_S(j)}\cdot\frac{W}{R_{b,S}^{\mathrm{UL}}(j)\cdot \nu_S^{\mathrm{UL}}(j)}} \qquad (1)$$

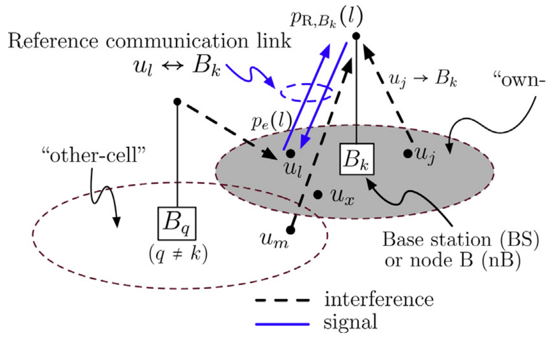


Fig. 2. Simplified representation of the communication signals (blue solid line) and interferences (black dashed lines) on the "reference" communication link $u_l \leftrightarrow B_k$. $p_{\mathrm{R},B_k}(l)$ represents the power received at the base station B_k emitted by user u_l. The total interference contains not only those interferences generated by the users assigned to the "own cell" ($u_j \in B_k$, $j \neq l$) but also those arising from other users located in "other cells" (user $u_m \in B_q$, $q \neq k$). Note that a user, like u_x, which is closer to the B_k, could be however assigned to another B_q. See the main text for further details. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

where:

- ξ is the ratio between the inter-interference (or "other-cell interference" [3], coming from users in other cells, B_q, $q \neq k$) and intra-interference (or "own-cell interference" [3]) produced by remaining users within the same own cell B_k. Usually ξ is assumed to be a constant average value ($\xi = 0.55$) in cells with omnidirectional antennas [3,53]. As will be shown later on, one of the novelties of our work consists in modeling ξ as a function that depends on the particular assignment of users to the base stations.
- $(e_b/n_0)_S(j)$ is the value for the ratio between the mean bit energy and the noise power density (including thermal noise and interference) required to achieve a given quality for service S. Note that this service could be different for each user u_j. For the purpose of this paper, $(e_b/n_0)_S(j)$ is an input parameter provided by the service requirements [3].
- $R_{b,S}^{\mathrm{UL}}(j)$ is the bit rate of service S in the j uplink within cell B_k. It is an input value stated by the service requirements. Throughout this paper, uppercases "UL" and "DL" will be used for labeling, respectively, uplink and downlink parameters.
- $\nu_S^{\mathrm{UL}}(j)$ is a utilization factor, which is 1 for data service, and $0 < \nu_S^{\mathrm{UL}}(j) < 1$ for voice services [3].

In a similar way, the downlink load factor in the cell served by nB B_k is [3]

$$\eta_{\mathrm{DL}}(B_k) \approx \left[(1-\bar{\alpha})+\bar{\xi}\right] \sum_{j=1}^{n_u^{B_k}} \frac{(e_b/n_0)_S(j)}{\frac{W}{R_{b,S}^{\mathrm{DL}}(j)\cdot \nu_S^{\mathrm{DL}}(j)}} \qquad (2)$$

where $\bar{\alpha}$ is an average orthogonality factor in the base station, and $\bar{\xi}$ is an average of ξ (across the cell), since in the DL, the ratio of other-base stations to own-base station interference depends on the user location and is thus different for each user j [3].

Finally, for the sake of clarity, Table 3 lists the symbols used in this paper.

4. Novel problem formulation

As mentioned in Section 1, one of the novelties of this paper when compared to the conventional approach and to our previous work [21] consists in modeling in a more accurate way the load factors (Section 4.1) and in constructing a cost function (with weights), which makes use of not only the load factors but also some ratios that measure the extent to which the other resources (aggregated capacity, power, codes) are used (Section 4.2). Those aspect related to the LHGGA algorithm we propose to tackled this problem will be explained in Section 5.

4.1. A more accurate model of interference and load factors

To clearly distinguish these from those stated by Expressions (1) and (2), and to ease the subsequent discussion, we label them as η_{UL}^{*} and η_{DL}^{*}, respectively.

We model the uplink load factor of cell B_k as

$$\eta_{\mathrm{UL}}^{*}(B_k) = \sum_{j=1}^{n_u^{B_k}} (1+\xi_{u_j \rightarrow B_k}^{\mathrm{UL}})\cdot \frac{1}{1+\frac{1}{(e_b/n_0)_S(j)}\cdot\frac{W}{R_{b,S}^{\mathrm{UL}}(j)\cdot \nu_S^{\mathrm{UL}}(j)}} \qquad (3)$$

where $\xi_{u_j \rightarrow B_k}^{\mathrm{UL}}$ is the ratio of other-cell to own-cell interference *on* the uplink communication (superscript "UL") between user u_j and node B_k, (subscript "$u_j \rightarrow B_k$"). As will be shown, $\xi_{u_j \rightarrow B_k}^{\mathrm{UL}}$ depends on the assignment to be done. If, as usual, the other-cell to own-cell interference ratio is assumed an average value ($\xi = 0.55$), then it

Table 3
List of symbols used in this work.

Symbol	Meaning
$\mathbf{A}_j$	Assignment vector of user u_j
$(e_b/n_0)_S(j)$	Ratio between the mean bit energy and the noise power density (including thermal noise and interference) required to achieve a given quality for service S of user number j.
$\bar{\alpha}$	Average orthogonality factor in the base station
B_k	Base station number k
$\mathcal{C}$	Cost function to be minimized
$\mathbf{c}_i$	Chromosome number i
$\mathcal{D}$	Statistical distribution
$\Delta_{n_u^{\mathrm{WS}}}$	Fraction of users without service
Δ_{Cod}	Fraction of codes
$\Delta_{C_{\mathrm{Ag}}^{\mathrm{DL}}}$	Fraction of aggregate capacity in downlink
$\Delta_{C_{\mathrm{Ag}}^{\mathrm{UL}}}$	Fraction of aggregate capacity in uplink
$\Delta_{P_{B_k}}$	Fraction of power emitted by base station B_k
$\eta_{\mathrm{DL}}(B_k)$	Downlink load factor
$\eta_{\mathrm{DL}}^{*}(B_k)$	Downlink load factor (proposed)
$\eta_{\mathrm{UL}}(B_k)$	Uplink load factor
$\eta_{\mathrm{UL}}^{*}(B_k)$	Uplink load factor (proposed)
$i_{u_j \to B_k}^{\mathrm{UL},B_q}$	Uplink interference (on link $u_j \to B_k$) generated by users assigned to other cells B_q
$i_{u_j \to B_k}^{\mathrm{UL},B_k}$	Uplink interference (on link $u_j \to B_k$) generated by users assigned to own cells B_k
ℓ_{u_m,B_X}	Total propagation loss in the link $u_m \to B_X$
L_{u_m,B_X}	Total propagation loss in the link $u_m \to B_X$, in dB
M	Number of base stations or nodes B
N	Number of users
n_u^{WS}	Number of user without service
$n_u^{B_k}$	Number of users assigned to base station (nodeB) B_k
$\nu_S^{\mathrm{DL}}(j)$	Utilization factor (of service S) in downlink j
$\nu_S^{\mathrm{UL}}(j)$	Utilization factor (of service S) in uplink j
$p_{\mathrm{R},B_k}(l)$	Power received at the base station B_k emitted by user u_l
$p_{e,u_m} \equiv p_e(m)$	Power emitted by user u_m
$p_{B_k \,\mathrm{\mid max}}$	Maximum power emitted by base station B_k
P_c	Crossover probability
P_m	Mutation probability
$\mathcal{P}_{\mathrm{size}}$	Population size
$R_{b,S}^{\mathrm{DL}}(j)$	Downlink bit rate of service S in the j downlink
$R_{b,S}^{\mathrm{UL}}(j)$	Uplink bit rate of service S in the j uplink
u_l	User l
w_η	Weight factor for load factors
$w_{\Delta_{C_A}}$	Weight factor of aggregated capacity ratio
$w_{\Delta_{P_{B_k}}}$	Weight factor of power ratio emitted by station B_k
$w_{\Delta_{\mathrm{Cod}}}$	Weight factor of code ratio
$w_{\Delta_{n_u}^{\mathrm{WS}}}$	Weight factor of fraction of users without service
W	Chip rate
ξ	Other-cell interference to own-cell interference ratio
$\bar{\xi}$	Average (across the cell) of other-cell interference to own-cell interference ratio
$\xi_{u_j \to B_k}^{\mathrm{UL}}$	Assignment-dependent other-cell interference to own-cell interference ratio (proposed)
$\Upsilon_{j,k}$	Signal to noise-and-interference ratio between u_j and all the base stations B_k

does not depend on index j and Expression (3) becomes into (1). To understand clearly why $\xi_{u_j \to B_k}^{\mathrm{UL}}$ depends on the assignment, it is first convenient to have a look at the way $\xi_{u_j \to B_k}^{\mathrm{UL}}$ is computed:

$$\xi_{u_j \to B_k}^{\mathrm{UL}} = \frac{i_{u_j \to B_k}^{\mathrm{UL},B_q}}{i_{u_j \to B_k}^{\mathrm{UL},B_k}} \tag{4}$$

where $i_{u_j \to B_k}^{\mathrm{UL},B_q}$ and $i_{u_j \to B_k}^{\mathrm{UL},B_k}$ are the uplink other-cell-interference (arising from users in other cells $B_q \neq B_k$) and UL own-cell-interference (from users inside the own cell B_k), respectively:

$$i_{u_j \to B_k}^{\mathrm{UL},B_q} = \sum_{u_m \,\in\, B_q,\, B_q \,\neq\, B_k} \frac{p_{e,u_m}}{\ell_{u_m,B_q}} \tag{5}$$

$$i_{u_j \to B_k}^{\mathrm{UL},B_k} = \sum_{u_m \,\in\, B_k,\, u_m \,\neq\, u_l} \frac{p_{e,u_m}}{\ell_{u_m,B_k}} \tag{6}$$

In the interference Expressions (5) and (6), p_{e,u_m} is the power that each user u_m emits, and ℓ_{u_m,B_X} is the total propagation loss in the link $u_m \to B_X$ (between user u_m and node B_X). Note that $B_X = B_k$ is the cell to be used for properly calculating the own-cell interferences while $B_X = B_q$, with $q \neq k$, the one for computing other-cell interferences.

A key point to note in Expressions (5) and (6), which is not easy to see intuitively, is that a generic user u_x, which can be physically located *inside* the cell coverer by its nearest base station (B_k in Fig. 2) could be however be *assigned* to another base station that is farther (B_q, $q \neq k$ in Fig. 2). We use the notation "$u_x \in B_q$" to mathematically express that user u_x has been assigned to nodeB B_q. If this is the case, $u_x \in B_q$ produces *other-cell interference* on the communication links involving B_k ($u_l \leftrightarrow B_k$).

That is, properly computing the different interference terms ($i_{u_j \to B_k}^{\mathrm{UL},B_k}$ and $i_{u_j \to B_k}^{\mathrm{UL},B_q}$) requires first to assign users to nBs. Only when the users have been assigned to nBs, by forming groups of users served by the assigned nBs, it is possible to know whether or not user u_m belongs to u_j' group (B_k) or to another, and to discern whether or not u_m in Expressions (5) and (6) produces other-cell interference (if $u_m \in B_q$, $q \neq k$) or own-cell interference (if $u_m \in B_k$). Depending on the way the assignment is done, $\xi_{u_j \to B_k}^{\mathrm{UL}}$ can have very different values.

In turn, the propagation losses in Expressions (5) and (6) can be computed as

$$\ell_{u_m,B_X} = 10^{L_{u_m,B_X}(\mathrm{dB})/10}, \tag{7}$$

$L_{u_m,B_X}(\mathrm{dB})$ being

$$L_{u_m,B_X}(\mathrm{dB}) = L_{u_m,B_X}^{\mathrm{PM}} + \sum_{\delta} L_\delta - \sum_{\tau} G_\tau, \tag{8}$$

where:

- $L_{u_m,B_X}^{\mathrm{PM}}$ can be computed by using a "propagation model" (PM) [55,56]. These propagation models for mobile communications are complex models, make use of many parameters (frequency, distance, antennas heights, and others [55–57]) and often need empirical adjustments: Cost231 Walfisch-Ikegami model [58] and Okumura–Hata propagation model [57,59–61]. For an urban macro cell with base station antenna height of 30 m, mobile antenna height of 1.5 m and carrier frequency f_C = 1950 MHz, the Okumura–Hata propagation model predicts a propagation loss

  $$L_{u_m,B_X}^{\mathrm{PM}}(\mathrm{dB}) = 137.4 + 35.2 \cdot \log\left[d(u_l, B_X)\right], \tag{9}$$

  $d(u_l, B_X)$ being the distance between the antennas of user's device u_l and base station B_X.
- $\sum_\delta L_\delta$ represent the remaining losses (body loss, cable loss in the base station, etc.)
- $\sum_\tau G_\tau$ is the sum of the antennas gains (both base station and user device).
- L_δ and G_τ are input data that depend on the service.

4.2. *Including more ratio parameters in the problem*

Each of these parameters aims to quantify the efficiency with which an available resource $\mathcal{R}$ is used. In this respect, the *utilization ratio* of resource $\mathcal{R}$ is defined ($\doteq$) as

$$\Delta_{\mathcal{R}} \doteq \frac{\mathcal{R}_{\mathrm{used}}}{\mathcal{R}_{\mathrm{available}}} \tag{10}$$

The first telecommunication resource whose use would be optimized is the available capacity for aggregating UL bit rates: $C_{\mathrm{Ag}}^{\mathrm{UL}}$. In any base station B_k, the UL bit rates of any user u_j for a service S, $R^{\text{UL}}_{b,S}(j)$, must be aggregated for ulterior backhauling. The corresponding aggregated capacity ratio in each B_k is defined as

$$\Delta_{C^{\text{UL}}_{\text{Ag}}} \doteq \frac{1}{C^{\text{UL}}_{\text{Ag}}} \sum_{j=1}^{n_u^{S_k}} R^{\text{UL}}_{b,S}(j), \tag{11}$$

Similarly, its counterpart for DL is defined as

$$\Delta_{C^{\text{DL}}_{\text{Ag}}} \doteq \frac{1}{C^{\text{DL}}_{\text{Ag}}} \sum_{j=1}^{n_u^{S_k}} R^{\text{DL}}_{b,S}(j) \tag{12}$$

Another important resource is the maximum power that the BS has in order to serve the active users. We model the efficiency in its use as

$$\Delta_{P_{B_k}} \doteq \frac{1}{p_{B_{k|\max}}} \sum_{j=1}^{n_u^{S_k}} p^{\text{DL}}_{B_k \to u_j} \tag{13}$$

where $p_{B_{k|\max}}$ is the available maximum power of base station B_k, and $p^{\text{DL}}_{e,B_k}(j)$ is the power emitted by B_k for serving this user j.

Finally, if N_S represents the number of different services ($N_S = 3$, in this study), and $N^{S_h}_{\text{Cod}}$ is the number of available codes for serving service S_h, the fraction on channelization codes in base station B_k is

$$\Delta_{\text{Cod}} \doteq \sum_{h=1}^{N_S} \frac{n^{B_k}_{u,S_h}}{N^{S_h}_{\text{Cod}}}, \tag{14}$$

Finally, the fraction of users *without* service is

$$\Delta_{n_u^{\text{WS}}} \doteq \frac{n_u^{\text{WS}}}{N} \tag{15}$$

where n_u^{WS} is the number of users without service.

These ratios, along with the load factors, will allow us to propose the novel cost function that we describe in the following subsection.

4.3. The complete mathematical formulation of the problem

Given a WCDMA network with M base stations and N active users, the problem consists in assigning (for each nB B_k, $k = 1, 2, \ldots, M$) the available resources (power, capacity and codes) to users by minimizing the cost function

$$\mathcal{C} = \frac{1}{M} \sum_{k=1}^{M} [w_\eta \cdot (\eta^*_{\text{UL}} + \eta^*_{\text{DL}}) + w_{\Delta_{C_A}} \cdot (\Delta_{C^{\text{UL}}_{\text{Ag}}} + \Delta_{C^{\text{DL}}_{\text{Ag}}} +)$$
$$+ w_{\Delta_{P_{B_k}}} \cdot \Delta_{P_{B_k}} + w_{\Delta_{\text{Cod}}} \cdot \Delta_{\text{Cod}} + w_{\Delta^{\text{WS}}_{n_u}} \cdot \Delta^{\text{WS}}_{n_u}], \tag{16}$$

constrained to the conditions that all the ratios (3)–(15) are real numbers ranging from 0 to 1. w_ϕ represents a weight factor for any of the involved components: load, utilization ratios of capacity, power, and codes, and fraction of users without services ($\phi = \eta_{\text{UL}}, \eta_{\text{DL}}, \Delta_{C^{\text{UL}}_{\text{Ag}}}, \Delta_{C^{\text{DL}}_{\text{Ag}}}, \Delta_{P_{B_k}}, \Delta_{\text{Cod}}, \Delta^{\text{WS}}_{n_u}$). Note that $0 \le \phi \le 1$ (see Expressions (3)–(15)). The weight values w_ϕ can be chosen depending on the importance we want to give to any of the physical quantities involved. This cost function differs from the one stated in [21] just in the introduction of these weight values.

To tackle this problem we propose the LHGGA with repair heuristics that follows.

5. Proposed Lamarckian Hybrid Grouping Genetic Algorithm with repair heuristics

The Lamarckian Hybrid Grouping Genetic Algorithm with repair heuristics we propose is a particular class of grouping genetic algorithm in which the unfeasible individuals are modified by a repair operator ("hybrid" GGA) and substituted by their corresponding repair version ("Lamarckian"). The grouping genetic algorithm is a class of evolutionary algorithm especially modified to tackle grouping problems, i.e., problems in which a number of items must be assigned to a set of predefined groups. In our problem, a number of N users have to be assigned to a number of M base stations. The grouping genetic algorithm was first proposed by Falkenauer [28,29], who realized that traditional genetic algorithms had difficulties when they were applied to grouping problems (basically, because the standard binary encoding increases the space search size in this kind of problems). In the GGA, the encoding, crossover and mutation operators of traditional genetic algorithms are modified to obtain a compact algorithm with very good performance in grouping problems, including telecommunication problems [31–34].

Assuming that the reader is familiar with the fundamentals the GGA is based on, the following sections focus on describing only the novel and/or the most particular aspects of the LHGGA we propose to tackle the difficulties underlying the problem stated in (4). Specifically, we emphasize the particular encoding of our problem (Section 5.1), the repair heuristic (Section 5.2), the selection operator (Section 5.3, and the crossover and mutation 5.4) operators.

5.1. Problem encoding

The encoding we use is a variation with respect to the classical grouping encoding proposed initially by Falkenauer [28,29]. In this classical approach, the encoding is based on separating each chromosome $\mathbf{c}$ into two parts: $\mathbf{c} = [\mathbf{e}|\mathbf{g}]$, the first one being the *element* section, while the second part, the *group* section. Since the number of base stations in our network is *constant* (M), we have used the following variations of the classical grouping encoding:

(1) The group section $\mathbf{g}$ is an $(M+1)$ length vector, whose elements (labeled $n_u^{B_j}$) represent the number of users assigned to each jth base station (B_j). Subscript j ranges from -1 to M, $j = -1$ being used to represent those users that are *not* connected to any node, that is, those in an "imaginary" or virtual base station that we have labeled "base station -1". As will be shown, this group part is necessary since the crossover operator acts on the group part and *not* on individual users. This is much more efficient than applying this operators on the element part (as in a conventional genetic algorithm) because, as proved by Falkenauer [28,29], classical genetic algorithms have difficulties in grouping problems because their encoding increases the space search.

(2) The element part $\mathbf{e}$ is an N-length vector whose elements ($u_j^{B_k}$) mean that user u_j has been assigned to base station B_k.

As an example, following our notation, in a trial solution with N elements (users) and M groups (base stations), a candidate assignment could be encoded by a chromosome

$$\mathbf{c}_i = [u_1^{B_h}, u_2^{B_p}, \ldots, u_i^{B_j}, \ldots, u_N^{B_w} \quad | \quad n_u^{B_{-1}}, n_u^{B_1}, n_u^{B_2}, \ldots, n_u^{B_j}, \ldots, n_u^{B_M}], \tag{17}$$

where $n_u^{B_{-1}}$ is the number of users without service (n_u^{WS}), those that have not been able to be assigned to any nB and do not have service. We represents this by assigning then to a "virtual" nB labeled B_{-1}.

Note that $\sum_{k=-1}^{M} n_u^{B_k} = N$, which simply states that the total number of active users in the network are distributed among the M base stations, including those $n_u^{B_{-1}} = n_u^{\mathrm{WS}}$ users which have not received telecommunication resources.

5.2. Chromosome repairing operator

The reason that compels us to propose a chromosome repairing operator, fully adapted to the problem at hand, is that the random generation of the initial population, or the crossover and mutation operators could produce candidate solutions which may have *no physical sense*. An example of such "bad candidates" could be a chromosome $\mathbf{c}_j$ encoding an individual in which one or several of the efficiency ratios (defined by (3)–(15) between 0 and 1) are however greater than 1. Intuitively, this is the case when $\mathbf{c}_j$ encodes, for instance, a particular assignment in which there is a base station with too many users that the aggregate DL capacity leads to a ratio $\Delta_{C_{\mathrm{Ag}}^{\mathrm{DL}}}(\mathbf{c}_j) > 1$. This is an "overload" in the sense that the algorithm is trying to use more capacity than the eNB actually has. Or in other words, a violation of one of the constrains our problem involves.

The proper management of the problem's constraint is a key point for obtaining "good-quality solutions". There are basically two groups of strategies aiming at properly managing problem constraints in evolutionary algorithms [22]: introducing *penalty terms* in the cost function [62–65], or *repair heuristics* [22,66–70]. We have selected repair heuristics in the detriment of penalty terms in the cost function because repair operators have been found to perform better in a most of Hybrid GGAs. See [22] for further details.

The propose chromosome repairing operator [22] works as follows:

(1) For any chromosome $\mathbf{c}_j$, compute its cost function, and check their constituent components $\phi(\mathbf{c}_j)$ to detect whether or not one or several overloads (thresholds violations) have arisen (for instance, $\Delta_{C_{\mathrm{Ag}}}^{\mathrm{DL}}(\mathbf{c}_j) > 1$). If not, the chromosome is keeping unchanged.
(2) But, if some overheads have been detected ($\phi(\mathbf{c}_j) > 1$), then the chromosome repairing operator selects at random a gene (user) in the chromosome element part, and assigns it to another eNB (also at random).
(3) The novel chromosome so generated is checked searching for overloads.
(4) This process is repeated until a maximum number of iterations is reached or the new assignment does not violate any constrain. If the process ends with the maximum number of iterations, this means that the user has not been able to be assigned to any BS without violating some constrain, and thus is assigned to virtual base station −1.

Note that as the unphysical chromosome is substituted by its repaired version after the application of the described repair heuristic, the algorithm is Lamarckian. This is why we have labeled our algorithm LHGGA.

5.3. Selection operator

Our selection operator is inspired by a rank-based wheel selection mechanism. In a first step, individuals are sorted in a list based on their quality (measured by their cost function ($\mathcal{C}(\mathbf{c}_i)$). The position of the individuals in the list is called *rank of the individual*, and are labeled R_i, $i = 1, \ldots, \mathcal{P}_{\mathrm{size}}$, $\mathcal{P}_{\mathrm{size}}$ being the population size. We consider a rank in which the best individual (lowest cost function) x is assigned $R_x = \mathcal{P}_{\mathrm{size}}$, the second best y, $R_y = \mathcal{P}_{\mathrm{size}} - 1$, and so on. Thus we can associate to each individual (assignment) i encoded by chromosome $\mathbf{c}_i$ a selection value

$$\Xi_i = \frac{2 \cdot R_i}{\mathcal{P}_{\mathrm{size}} \cdot (\mathcal{P}_{\mathrm{size}} + 1)} \qquad (18)$$

Note that these values Ξ_i, $i = 1, \ldots, \mathcal{P}_{\mathrm{size}}$, are *normalized* between 0 and 1, depending on the position of the individual in the ranking list. It is worth emphasizing that this rank-based selection mechanism is static, in the sense that *probabilities of survival* (given by Ξ_i) do not depend on the generation, but on the position of the individual in the list.

The process carried out by our algorithm consists in selecting the parents for crossover using this selection mechanism. This process is performed with replacement, i.e., a given individual can be selected several times as one of the parents, however, individuals in the crossover operator must be different. The final number of individuals that will be replaced by those obtained with the crossover operator depends on a *crossover probability*, fixed in 80% of the individuals for a given generation of the algorithm. This and other details of the tailored crossover and mutation operators we propose are described in the following section.

5.4. Crossover and mutation operators

Note that, because of the particular encoding of the problem used in this work (Section 5.1), the number of groups ($M+1$, because one of them is used to store the number of users which have not been assigned to a base station) in each individual of the GGA is *fixed* (since the number of base stations M is a constant parameter in the problem at hand). Due to this peculiarity, we propose a novel crossover procedure fully adapted to the problem at hand. It works in a two-parents one-offspring fashion, in the following way:

(1) Randomly select two individuals for the crossover operation (father and mother).
(2) Generate the initial offspring as a simple copy of the father.
(3) Choose randomly K groups from non-empty groups in the father.
(4) The users that were assigned to these K groups in the mother individual, are now re-allocated to the corresponding groups in the offspring.

To assist in understanding this, Fig. 3 shows an example of the crossover implemented in the GGA. It is a simple case with $N = 10$ users and $M = 3$ base stations. For the sake of clarity, only 1 non-empty group ($K = 1$) in the father has been selected: in this case, the third group (represented inside a dashed square), which corresponds to base station 2, B_2. This is because, as stated in our encoding (17), the *first* group is used to store the number of users without service (in this case, zero users). To help understand this example, we remark the considerations that follows.

- The first considerations are related to the chromosome *group part*. For the sake of clarity, let us focus on the father represented in the uppermost part of Fig. 3. The first position of its group part, used to quantify the number of users without service, is 0, what means that *all* users have been assigned to a base station. The remaining three positions (because $M = 3$ base stations) are used to store the number of users $n_u^{B_k}$ assigned to each base station B_k. Specifically, the third position (dashed square) in the father's group part is used to store the number of users (3, in this example) in base station B_2 ($\Rightarrow n_u^{B_2} = 3$). Note that it is 3 because there are 3 users assigned to base station B_2 (those represented into boxes in the element part of the father: users at positions 1, 3 and 10, respectively). To clearly understand the meaning of the

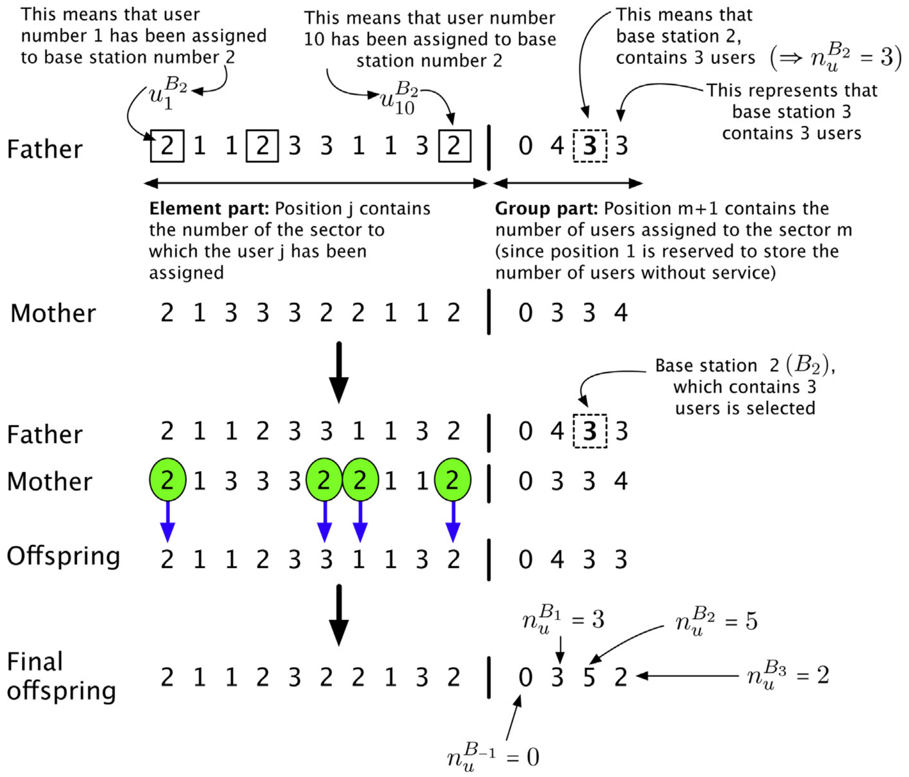

This means that user number 1 has been assigned to base station number 2
This means that user number 10 has been assigned to base station number 2
This means that base station 2, contains 3 users $(\Rightarrow n_u^{B_2} = 3)$
$u_1^{B_2}$
$u_{10}^{B_2}$
This represents that base station 3 contains 3 users
Father 2 1 1 2 3 3 1 1 3 2 | 0 4 3 3
Element part: Position j contains the number of the sector to which the user j has been assigned
Group part: Position m+1 contains the number of users assigned to the sector m (since position 1 is reserved to store the number of users without service)
Mother 2 1 3 3 3 2 2 1 1 2 | 0 3 3 4
Base station 2 (B_2), which contains 3 users is selected
Father 2 1 1 2 3 3 1 1 3 2 | 0 4 3 3
Mother 2 1 3 3 3 2 2 1 1 2 | 0 3 3 4
Offspring 2 1 1 2 3 3 1 1 3 2 | 0 4 3 3
Final offspring 2 1 1 2 3 2 2 1 3 2 | 0 3 5 2
$n_u^{B_1} = 3$
$n_u^{B_2} = 5$
$n_u^{B_3} = 2$
$n_u^{B_{-1}} = 0$

Fig. 3. Outline of the tailored crossover operator used in the proposed LHGGA. In this example, the number of users to be assigned is $N = 10$, and the number of base stations is $M = 3$, what makes the number of groups be $M + 1 = 4$. See the main text for further details.

element part, it is convenient to remember that position j contains the number of the base station to which the user j has been assigned. For instance, position 1 contains "2", what means that user number 1 has been assigned to base station number 2: $u_1^{B_2}$, following the notation we first stated in Section 4.

- As mention, in this example, only $K = 1$ non-empty group in the father has been randomly selected: the third group inside the dashed squared, which corresponds to base station 2. This implies that the mother's elements assigned to base station 2 (inside circles) have to be copied into their corresponding positions in the offspring, as shown with the blue arrows. This generates a final offspring in which there are 3 users in base station B_1 (by simply counting the number of "1" on the element part $\Rightarrow n_u^{B_1} = 3$), 5 users in B_2 $(\Rightarrow n_u^{B_2} = 5)$, and 2 user in B_3 $(\Rightarrow n_u^{B_3} = 2)$. Note that $n_u^{B_1} + n_u^{B_2} + n_u^{B_3} = 3 + 5 + 2 = 10 = N$ users. That is, all users have been assigned to a base station. This implies that the number of users with no service is $(0 \Rightarrow n_u^{B_{-1}} = 0)$ and, consequently, the first position in the group part is "0". This is why its group part is 0 3 5 2.

Note that the number of groups in the offspring individual still remains fixed to $M + 1$, and that, in general, some of the groups could become empty because of the crossover operation.

Regarding the mutation operator, we perform a per-gene mutation on the users region of the chromosomes, and each user selected to be mutated is re-assigned to a region different to its current region.

6. Experimental work

We have organized this section as follows:

- For comparative purposes, Section 6.1 summarizes our implementation of the conventional approach that aims to minimize only the load factors.
- Aiming at having a unified framework for all the experiments, in which these may be replicated by other researchers, Section 6.2 focuses on describing the layout of the base stations that we have considered. In this fixed deployment we have studied different scenarios in which we vary the number of users, their distributions, or service profiles (Sections 6.4–6.6).
- Once the experimental setup has been set, Section 6.3 compares the performance of the proposed LHGGA to that of the conventional approach (CA) stated in Section 6.1.
- Section 6.4 makes use of the proposed LHGGA to explore the influence of the number of users: $N = 250$, 500, 750 active users.
- Section 6.5 applies the proposed LHGGA to study the influence of different distributions of users, for instance, to model the situation in which users tend to concentrate in a cell because of an unexpected event.
- Finally, Section 6.6 focuses on the applicability of the LHGGA to different service profiles, and studies to what extent the number of users without service increases as the percentage of users with data services rises in the detriment of the voice service.

6.1. Conventional approach

Based on the background summarized in Sections 2 and 3, in the conventional approach, a user u_j is assigned to a node B_k only if the increment in the loads and interferences do not violate some predefined thresholds [53,71,72]. The problem of assigning users to base stations can be tackled in a conventional approach by using classical cell selection algorithms such as BSCS or RPCS [18,3] (which were reviewed in Section 2) or variations of them.

Specifically, we have considered the following combination of the BSCS and RPCS algorithms. For any user u_j (with $j = 1, 2, \ldots, N$), we compute the SINR between u_j and all the base stations B_k (with $k = 1, 2, \ldots, M$): $\Upsilon_{j,k}$. This leads to a $N \times M$ matrix $\left[\Upsilon_{j,k}\right]_{N\times M}$ of SINR ratios. For any user u_j, we compute an "assignment vector",

Table 4
Values of the services parameters. ARM means Adaptive Multi-Rate.

Service, S_i	$(E_b/N_0)_i$ (dB)	$R^{\text{UL}}_{b,i}$ (kbps)	$R^{\text{DL}}_{b,i}$ (kbps)	$\nu^{\text{UL}}_i = \nu^{\text{DL}}_i$	$N^{\text{DL}}_{C_i}$ (codes)
"1" (ARM voice)	5	12.2	12.2	0.58	256
"2" (data)	1.5	64	64	1	32
"3" (data)	1	64	384	1	4

$\mathbf{A}_j$, which contains a list of BSs, sorted from the one that provides the best SINR to the one that gives the worst one. Initially, each user u_j is assigned to the nodeB with the corresponding best SINR ("best base station" (BBS) in the BSCS algorithm [3,18]), that is, to the first one of the assignment vector $\mathbf{A}_j$. In any cell, the algorithm checks whether or not the assignment leads to a load factor higher that the threshold [53] (overload). In each overloaded cell (let say, for instance, B_g), the user with the worst SINR with respect to B_g (let say, for instance, u_f) is detached from B_g and assigned to the next non-overloaded BS of its assignment vector $\mathbf{A}_f$. The algorithm iterates until either the cells are no longer overloaded, which may cause some users fail to be assigned to any station.

Like the BSCS algorithm [18,19], this algorithm leads to an efficient use of radio resources, but suffers from inefficiencies because the aggregate capacity of the BBS could be saturated ("overloaded").

6.2. Base station layout and experimental set up

6.2.1. Services and service profiles

We have considered three different services, labeled S_i = "1", "2", and "3" in Table 4: while S_1 is an Adaptive Multi-Rate (ARM) voice service with $R^{\text{DL}}_{b,1} = 12.2$ kbps, S_2 and S_3 are data services at $R^{\text{DL}}_{b,2} = 64$ kbps and $R^{\text{DL}}_{b,3} = 384$ kbps, respectively. The characteristic values of the parameters used for these three services have been listed in Table 4, these parameters being: $(E_b/N_0)_i$ (dB), $R^{\text{UL}}_{b,i}$ (kbps), $R^{\text{DL}}_{b,i}$ (kbps), $\nu^{\text{UL}}_i = \nu^{\text{DL}}_i$, and $N^{S_h}_{\text{Cod}}$ (codes).

With these services, there may be different "service profiles". In this respect, for the sake of clarity, and in the effort of testing our algorithm, we have considered two different service profiles. "Service Profile" 1 (SP1) has 90% of users with service S_1, 9% with S_2 and 1% with S_3. "Service Profile 2" (SP2) is used to explore the influence of an increasing number of users with data service in the detriment of voice service: 67% of users with service S_1, 24% with S_2 and 9% with S_3.

6.2.2. Base stations layout and network parameters

Fig. 4 shows the layout of the base stations that we have considered in this experimental work. Its purpose is to have a unified framework for all the experiments, in which these may be replicated by other researchers. The base station layout represented in Fig. 4 is a deployment with $M = 9$ base stations (red circles) distributed in a 16 km^2 area. The minimum distance between two base stations is 1 km. For the sake of clarity, the number of users represented in Fig. 4 is $N = 100$. These users are randomly distributed with a uniform distribution. We label it as D1 to clearly distinguish it from other distribution (D2), which will be described in Section 6.5. Black +, green *, and blue $\circ$ symbols represent the users with service S_1, S_2 and S_3, respectively.

It is worth mentioning, on the one hand, that although in Fig. 4 we have considered a case with $N = 100$ users for clarity, the influence of a variable number of users ($N = 250, 500, 750$) have also been studied in Section 6.4. On the other hand, although in Fig. 4 the location of the N users have been randomly generated with a uniform distribution (D1), nonetheless, Section 6.5 explores the influence of other statistical distribution D2, like the one represented in Fig. 5. For illustrative purposes, we have considered in 5 $N = 100$ users, in which 50% of users have a Gaussian distribution (with mean value ($x = 1$ km, $y = 1$ km) – i.e., around nB B_1 – and standard deviation 1 km) and the remaining 50% are distributed uniformly on the rest of the area.

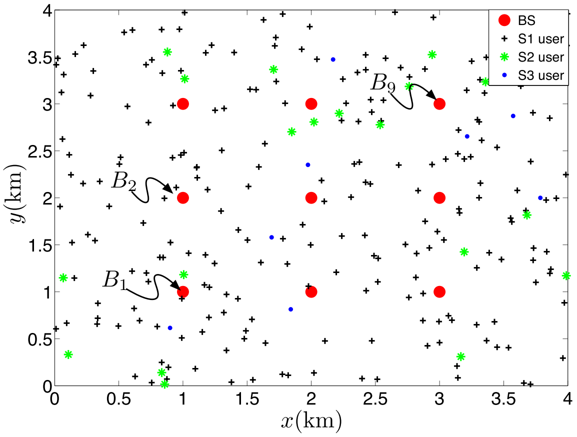


Fig. 4. Example of a deployment with $M = 9$ base stations (red circles). B_k (with $k = 1, 2, \ldots, 9$) labels the base stations (nodes B). The minimum distance between two base stations is 1 km. Black +, green *, and blue $\circ$ symbols represent the users with service S_1, S_2 and S_3, respectively. For the sake of clarity, the number of users is $N = 100$, randomly distributed with a uniform distribution. In other examples, N and its statistical distribution can adopt other values. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Finally, apart from those parameters related to the classes of service, there are other numerical data for the network parameters used in our experiments [3]: $\bar{\alpha} = 0.65$, $\bar{\xi} = 0.55$, $W = 3.84$ Mchip/s, $p_{B_{k|\max}} = 36$ W, and $C^{\text{UL}}_{\text{Ag}} = C^{\text{DL}}_{\text{Ag}} = 1536$ kbps.

6.2.3. LHGGA parameters

The values of the LHGGA parameters that have been found to work well in our experimental work are: crossover probability $P_c = 0.8$, mutation probability $P_m = 0.01$, and population size $\mathcal{P}_{\text{size}} = 500$ individuals.

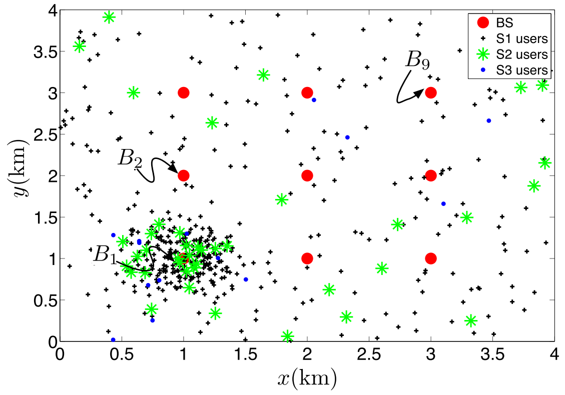


Fig. 5. Example of a deployment with $M = 9$ base stations (red circles) and $N = 500$ users with a non-uniform distribution. 50% of users have a Gaussian distribution with mean value ($x = 1$ km, $y = 1$ km) – i.e., around nB B_1 – and standard deviation 1 km. Black +, green *, and blue $\circ$ symbols represent the users with service S_1, S_2 and S_3, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

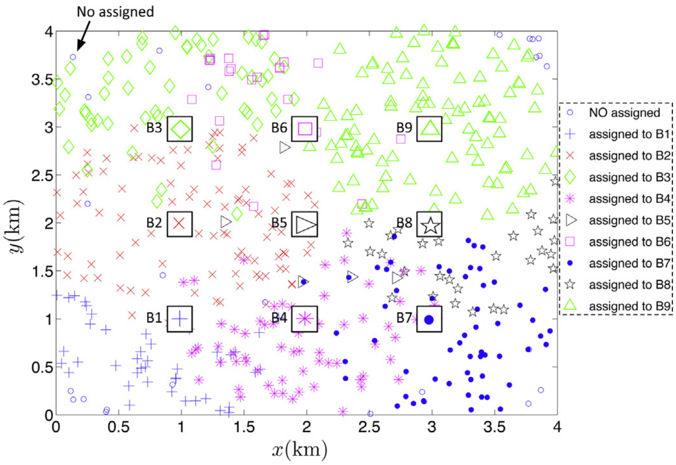


Fig. 6. Assignment of $N = 500$ uniformly distributed users to $M = 9$ base stations achieved by the LHGGA algorithm. Each BS has been represented with a different symbol (+, ×, ◊, ▷, ···), so that any user attached, for instance, to B_3 (◊-symbol), has been represented with that symbol (◊).

We have carried out 20 runs of each LHGGA algorithm, with 300 generations each. This generation number has been found to be large enough for the algorithm to converge.

Once we have completed the description of the experimental setup, we can begin now to compare our method with the conventional one.

6.3. Comparing LHGGA and the conventional approach

Figs. 6 and 7, which show respectively the different assignments that the LHGGA and CA algorithms have found, will assist us in discussing this issue. Each BS in Figs. 6 and 7 has been represented by a square box containing a different symbol (+, ×, ◊, ▷, ···), so that any user attached, for instance, to base station B_3 (◊-symbol inside the box), will be represented with that symbol (◊). This notation helps properly analyze and understand the different assignments that the LHGGA and CA algorithms generate. They correspond to $N = 500$ users, which leads to a user density $D_U \approx 31.25$ users/km^2.

Note that both Figs. 6 and 7 have *identical* user locations, but *differ in the way they are assigned to different stations*. This can be easily seen by taking a look at those users located in-between base stations B_6 and B_9 in both figures. While in Fig. 6 the users are mostly labeled with green △-symbols (what means that they have been assigned to B_9 – △ symbol –), however, in Fig. 7, many of these users located between stations B_6 and B_9 (which in Fig. 6 were mostly assigned to B_9) are now however *without service* (represented with blue ○ symbols) since they have not been assigned to any nB. The conventional approach assignment (Fig. 7) works worse in the sense that it leaves more customers unserved.

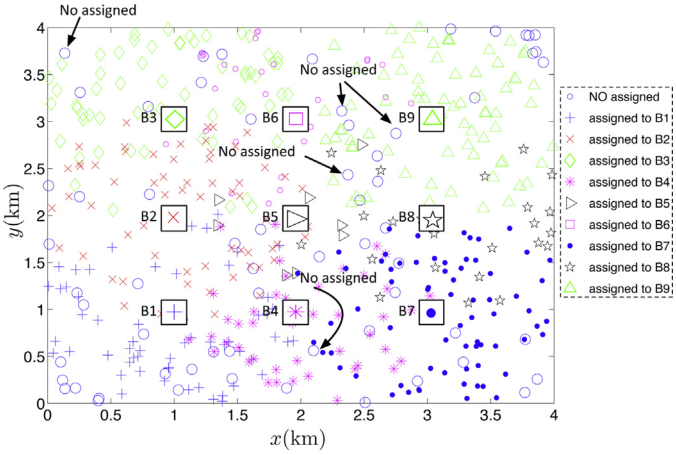


Fig. 7. Assignment of $N = 500$ uniformly distributed users to $M = 9$ base stations achieved by the CA algorithm. Like in Fig. 6, any BS has been represented with a different symbol (+, ×, ◊, ▷, ···), so that, for instance, any user assigned to B_3 (◊-symbol) has been represented with that symbol (◊).

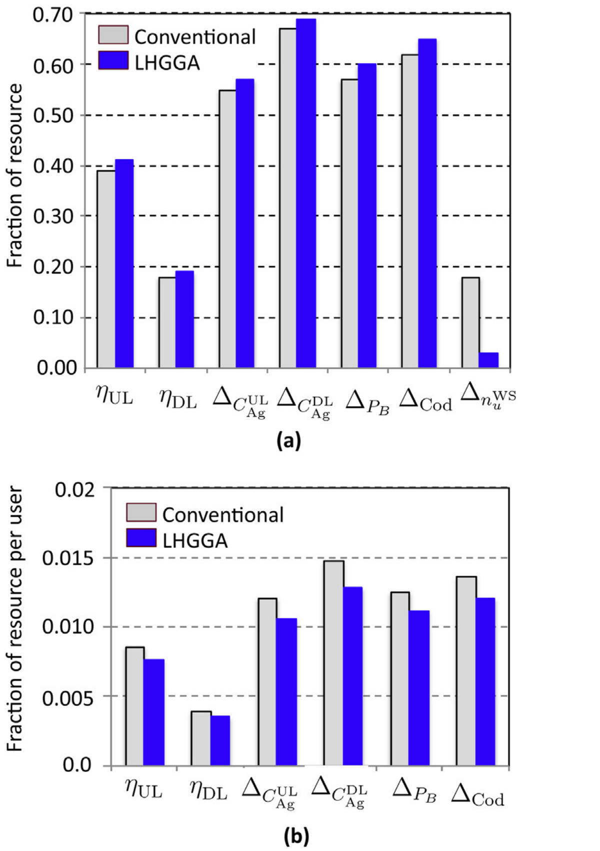


Fig. 8. (a) Cost function constituent elements corresponding to the user assignment computed by the conventional approach (gray bars) and the LHGGA method (blue bars). (b) Fraction of resources *per* user found by the conventional approach (gray bars) and the LHGGA method (blue bars). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

To proceed further in this regard it is convenient to focus on Fig. 8(a). It compares, respectively, the value of the different constituents ($\phi_\varsigma = \eta_{\mathrm{UL}}, \eta_{\mathrm{DL}}, \Delta_{C_{\mathrm{Ag}}^{\mathrm{UL}}}, \Delta_{C_{\mathrm{Ag}}^{\mathrm{DL}}}, \Delta_{P_{B_k}}, \Delta_{\mathrm{Cod}}, \Delta_{n_u}^{\mathrm{WS}}$) of the minimized cost function, computed by the conventional approach (gray bars) and by the proposed LHGGA method (blue bars).

The most relevant aspect arising in Fig. 8(a) is that the LHGGA method assigns resources to *many more users*: the fraction of users *without service* in the LHGGA assignment is only $\Delta_{n_u}^{\mathrm{WS}}|_{\mathrm{LHGGA}} = 3\%$ (mean value with standard deviation 1.2×10^{-4} over 20 runs). This represents only 15 users in absolute terms, which is much smaller than that achieved by the conventional assignment, which is $\Delta_{n_u}^{\mathrm{WS}}|_{\mathrm{CA}} = 18\%$ (i.e., 90 users). Note that $\Delta_{n_u}^{\mathrm{WS}}|_{\mathrm{LHGGA}}$ is 6 times smaller than $\Delta_{n_u}^{\mathrm{WS}}|_{\mathrm{CA}}$. In this respect, the LHGGA strategy is more practical for the operator economical strategy since it help increase the number of active users without having to draw upon novel and cost deployments. Additionally, the higher service availability is perceived by users positively and help operator increase market share.

However, a critical look at Fig. 8(a) could lead to the misleading conclusion that the LHGGA algorithm does not minimize the remaining parameters as well as the CA does. This could arise from the observation that the values adopted by the remaining parameters ($\eta_{\rm UL}$, $\eta_{\rm DL}$, $\Delta_{C_{\rm Ag}^{\rm UL}}$, $\Delta_{C_{\rm Ag}^{\rm DL}}$, $\Delta_{P_{B_k}}$, $\Delta_{\rm Cod}$) computed by the CA are slightly lower that those provided by the LHGGA approach. This is because, as the CA has only been able to assign resources to a reduced amount of users ($500 - 90 = 410 = N_{\rm u}^{\rm CA}$), then they use (as a whole) a resource fraction smaller than that of the $500 - 15 = 485 = N_{\rm u}^{\rm LHGGA}$ users that our LHGGA has properly assigned. The fraction of resources used by the $N_{\rm u}^{\rm LHGGA}$ users is slightly superior (in absolute terms). This makes sense because there are more active users (communication links) that, as a whole, consume more resources in absolute terms.

However, in *relative terms* (that is, in resources used *per* active communication link – or active user –), the situation is quite different. This has been illustrated in Fig. 8(b), in which the resources assigned have been *normalized* by the number of *served users*. Note that the proposed method leads to an assignment in which there are more customers with the required service along with a *lower consumption-per-user* than that achieved by the CA. The mean value of the resources consumed per user in the LHGGA assignment is 88.40% of that consumed by users assigned by the CA.

The key conclusion deduced from this experiment is that the proposed LHGGA exhibits a *superior* performance than that of the conventional method (which minimizes only the load factors). The proposed LHGGA not only assigns resources to more users (97% of users, higher than 82% of users assigned by the CA) but also it does it more efficiently, since the mean value of the resources consumed per user in the LHGGA assignment is 11.60% lower than that of the CA assignment.

Once we have illustrated the good performance of the LHGAA proposal, the following sections aim at checking its performance in more complex scenarios. Let's start studying a scenario in which the number of users varies.

6.4. Influence of number of users

In these experiments with varying number of users, we have considered, as above, that these users are distributed in a uniform way (distribution D1). The effects arising when users are located using other distribution will be postpone to Section 6.5.

Specifically, we have explored a scenario with three different numbers of users, $N = 250$, 500, 750 users, which lead to user densities $D_{\rm U} \approx 15.62$, 31.25, 46.87 users/km^2/frequency, receptively.

In turn, we have carried out two classes of experiments. They are related to the fact, mentioned before, that our cost function is flexible in the sense it contains weights that assist the engineer to emphasize a parameter or another. To illustrate this potential we have considered here two classes of experiments with increasing number of users. The first one consists in using the cost function to be minimized without imposing any restrictions on the weights (that is, all are unity), and explores the influence of the increasing number of users in the network. The second set of experiments consists in using the cost function with weights that help the LHGGA reach an assignment that minimizes the number of users without service.

In this respect, Figs. 9 and 10 will assist us in exploring these effects. Fig. 9 represents the mean value and standard deviation (over 20 runs of the LHGGA) of the cost function components computed by the LHGGA method as a function of the number of users $N = 250$ (blue bars), 500 (red bars), and 750 (green bars) users. The cost function has *no* restriction on their weights (that is, $w_\eta = w_{\Delta_{C_A}} = w_{\Delta_{P_B}} = w_{\Delta_{\rm Cod}} = w_{n_u^{\rm WS}} = 1$). On the contrary, the cost function minimized in Fig. 10 corresponds to $w_\eta = w_{\Delta_{C_A}} = w_{\Delta_{P_{B_k}}} = w_{\Delta_{\rm Cod}} = 1$ and $w_{n_u^{\rm WS}} = 10$.

Comparing Figs. 9 and 10 is easy to note that using the cost function without imposing any restrictions on the weights (that is, $w_\phi = 1, \forall\phi$), the algorithm tends to minimize the components as a whole (Fig. 9), while using $w_{n_u^{\rm WS}} = 10$ (Fig. 10) leads to better solutions in terms of network availability (since there are far fewer users without service). The designer has the freedom to select the most suitable weights to the interests of the mobile operator.

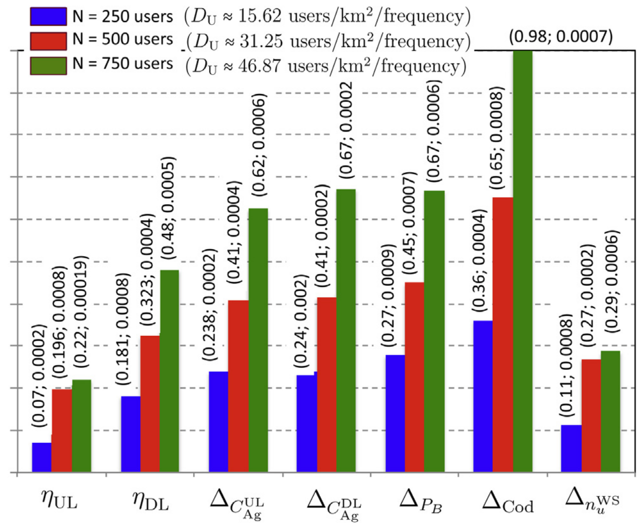


Fig. 9. Mean value and standard deviation of the cost function constituent elements corresponding to the user assignment computed by the proposed LHGGA as a function of the number of users $N = 250$, 500, 750 users. The cost function has weights $w_\phi = 1$.

6.5. Influence of different users distributions

Aiming at testing the algorithm in a common situation in which part of the customers tend to be concentrated in a cell, we have designed the following experiment: the $N = 500$ users are randomly distributed so that 50% are distributed around base station B_1 with a Gaussian distribution (mean = (1, 1) km, standard deviation = 1 km), and the other half of users are distributed uniformly on the rest of

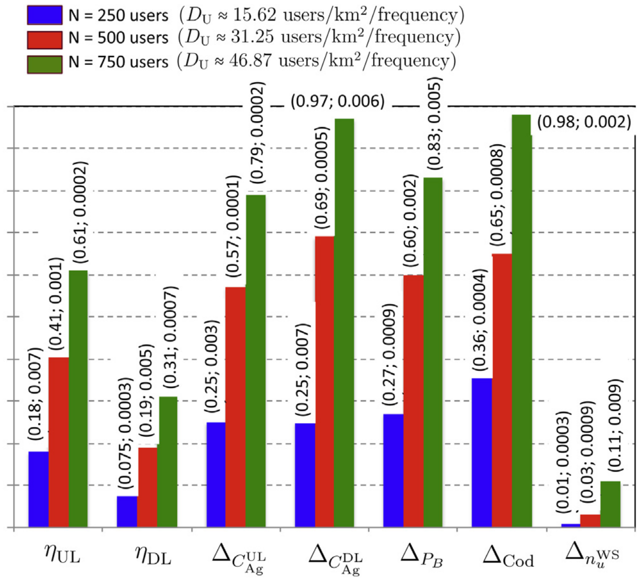


Fig. 10. Mean value and standard deviation of the cost function constituent elements corresponding to the user assignment computed by the proposed LHGGA as a function of the number of users $N = 250$, 500, 750 users. The cost function has a weight $w_{n_u^{\rm WS}} = 10$, while the remaining are unity.

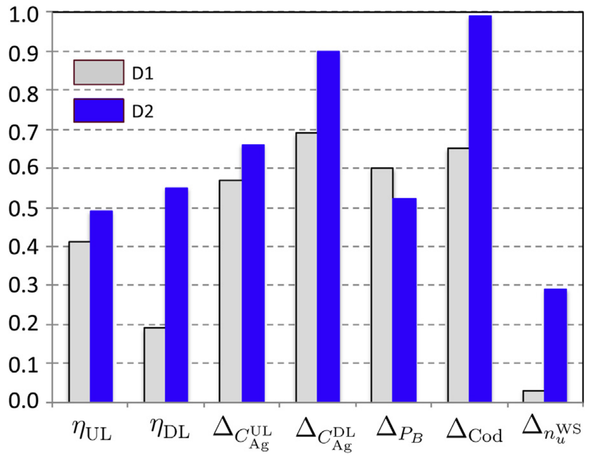


Fig. 11. Results computed by LHGGA as a function of two different distribution of the $N = 500$ users with service profile SP1. D1 corresponds to a uniform distribution. D2 is a distribution in which 50% of user are distributed around B_1 (mean = (1, 1) km, standard distribution = 1 km) and the other 50% of user are distributed uniformly on the rest of the 4 km × 4 km area.

the area. He have labeled this distribution “D2” to distinguish it from the uniform distribution we have used so far (D1).

Fig. 11 shows the results computed by the LHGGA algorithm as a function of the two aforementioned distributions (gray bar: distribution D1; blue bars: distribution D2). When comparing to the results corresponding to the uniform distribution used hitherto (D1), an interesting point to note is that the fraction of users without service increases from 0.03 (in distribution D1) to 0.29 (D2). In the same trend, the fraction of codes used approaches 0.99, which probably leads to leave many users without service, as shown in Fig. 12. It represents, for clarity, only those users without service. Note that most of them are in the area where there is a higher concentration due to the Gaussian distribution around B_1. The results represented in Fig. 11 make sense since the higher user density in cell B_1 has saturate the number of available codes, leading to the increased number of users without service.

6.6. *Influence of service profiles*

To explore the influence of an increasing percentage of users demanding data service in the detriment of the voice service, we have considered two different service profiles. “Service Profile 1” (SP1) has 90% of users with service S_1, 9% with S_2 and 1% with S_3. Service Profile 2 (SP2) is used to explore the influence of an increasing number of users with data service in the detriment of voice service: 67% of users with service S_1, 24% with S_2 and 9% with S_3.

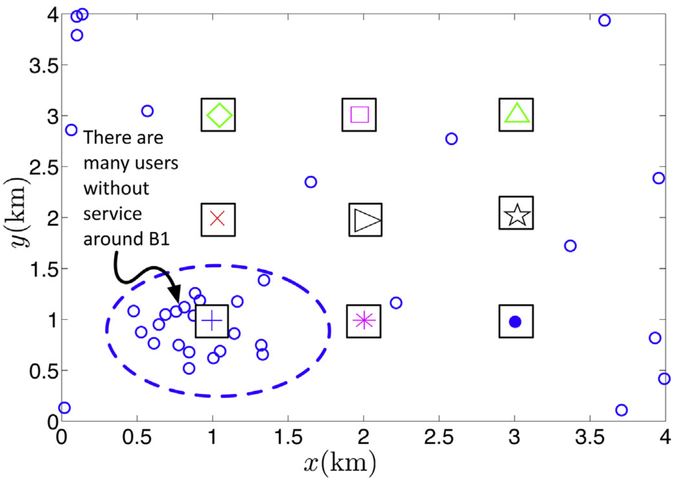


Fig. 12. Location of users without service in the assignment of Fig. 11. Most of them are in the area where there is a higher concentration due to the Gaussian distribution around B_1.

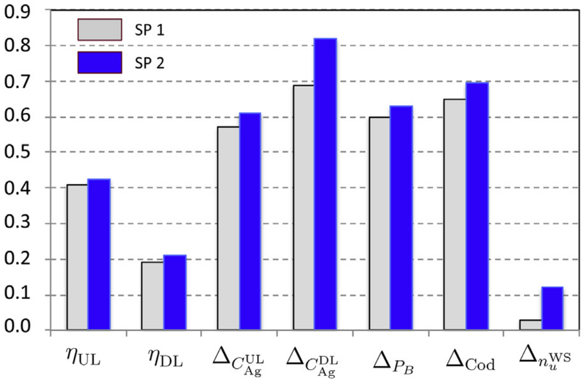


Fig. 13. Results obtained by the LHGGA for two different service profiles. Service profile SP1 corresponds to 450 users with service S_1, 45 with S_2, and 5 with S_3. Service profile SP2 corresponds to 335 users with S_1, 120 with S_2, and 45 with S_3.

Fig. 13 shows the results obtained for these two different data service penetration. Service profile SP1, which is the one that we have used in all the previous experiments, corresponds (for $N = 500$ users) to 450 users with service S_1, 45 with S_2, and 5 with S_3, while case SP2 corresponds to 335 users with S_1, 120 with S_2, and 45 with S_3. Note that the fraction of users without service increases from 0.03 (SP1, gray bar) up to 0.12 (SP2). The same trend is observed in the DL capacity (since the data services requires more kbps than the voice service) and in the fraction of codes. This make sense since a greater percentage of user with data service may lead to use more intensively the limited number of codes for data services, which are much less than those for voice (see Table 4).

7. Summary and conclusions

In this work we have tackled the problem of assigning resources (channelization codes, aggregated capacity, power) of M base stations (nodes B) to N users in Wide-band Code Division Multiple Access (WCDMA) networks by proposing two sets of novelties. The first group is related to the problem formulation itself while the second one focuses on tackling such problem formulation by using a Lamarckian Hybrid Grouping Genetic Algorithm (LHGGA) with repair heuristics.

The first contribution to the problem formulation consists in modeling in detail all the interference terms on any communication link. This formulation is different from the GGA approach [21], which made used of *approximated* expressions of the load factors (with inferences arriving from other cells modeled as an average value, an approximation that is often used when dimensioning WCDMA networks [13–17]). Considering all the interferences is crucial because these strongly depend on whether any user is assigned to a base station or other. The second contribution to the problem formulation is the proposal of a cost function (to be minimized) whose constituent elements (load factors, fractions of available resources – aggregated capacity, power, codes – and fraction of users without service) are multiplied by weight factors. This helps prioritize one or several constituents, for instance, in the effort of reducing the number of users without user, an ongoing concern for mobile operators. This cost function is different from that in [21] in which all the constituents had the same contribution.

The second group of novelties focuses on tackling this constraint problem (since all the constituent elements ϕ in the cost function must be $\phi \leq 1$) by using an LHGGA with repair heuristics to manage the generation of unfeasible chromosomes (encoding trial solutions with no physical meaning $\phi > 1$). A GGA is proposed because its crossover operator is able to acts on groups (users assigned to a base station) in an efficient way, well established in literature. The proposed GGA is called hybrid because the chromosome repairing operator is just used as a constraint-handling algorithm to shrink the search space to only individuals with physical meaning, and called Lamarckian because the unfeasible chromosome is substituted by its repaired version. We have defined a novel encoding scheme specific for the problem at hand, and we have also proposed different variations of the GGA crossover and mutation operators, suited for assignment problems in WCDMA networks.

The first conclusion deduced from our experimental work is that the proposed LHGGA exhibits a *superior* performance than that of the conventional method (which minimizes only the load factors). Specifically, the proposed LHGGA assigns resources to more users (97% of users in a scenario with 31.25 users/km^2, uniformly distributed) than the conventional one (82% of users), along with a reduction of the resources-per-user: the mean value of the resources consumed per user in the LHGGA assignment is 88.40% of that consumed by users assigned by the conventional one.

The LHGGA algorithm has been tested in a variety of scenarios.

We have explored the influence of different user concentrations in a 4 km×4 km area. The LHGGA algorithm has been found to be able to assign resources to the $M = 9$ base stations, with only 1% of users without service (for $N = 250$ users, 15.62 user/km^2/frequency), 3% (for $N = 500$ users, 31.25 user/km^2/frequency), and 11% (for $N = 750$ users, 46.87 user/km^2/frequency).

We have also tested the algorithm in a common situation in which part of the customers tend to be concentrated in a cell, 50% distributed around node B_1 with a Gaussian distribution (mean = (1,1) km, standard deviation = 1 km), the other half of users being uniformly distributed on the rest of the area. When comparing to the results corresponding to the uniform distribution, an interesting point to note is that the fraction of users without service increases from 0.03 to 0.29. In the same trend, the fraction of codes used approaches 0.99, which leads to leave many users without service, most of them in the area where there is a higher concentration. The results make sense since the higher user density around B_1 make use of all the available codes of such base station, leading to the increased number of users without service.

Finally, to explore the influence of an increasing percentage of users demanding data services (S_2 and S_3) in the detriment of voice service (S_1), we have considered two different service profiles. "Service Profile 1" (SP1) has 90% of users with service S_1, 9% with S_2 and 1% with S_3. Service Profile 2 (SP2) is used to explore the influence of an increasing number of users with data service in the detriment of voice service: 67% of users with service S_1, 24% with S_2 and 9% with S_3. The algorithm predicts that the fraction of users without service increases from 0.03 (SP1) up to 0.12 (SP2). The same trend is observed in the DL capacity (since the data services require more kbps than the voice service) and in the fraction of codes. This make sense since a greater percentage of user with data service may lead to use more intensively the limited number of codes for data services.

This second set of experiments lead to the final conclusion that the proposed algorithm predicts well all phenomena that are well known empirically by mobile operators.

Acknowledgments

This work has been partially supported by the project TIN2014-54583-C2-2-R of the Spanish Ministerial Commission of Science and Technology (MICYT), and by the Comunidad Autónoma de Madrid, under project number S2013ICE-2933_02.

References

[1] HSPA Operator Commitments report: 582 launched, incl. 182 845 42 Mbs DC-HSPA+ and_rst 63 Mbps 3C-HSPA+. Information Paper available at http://www.gsacom.com (July 2015).

[2] T. Chapman, E. Larsson, P. von Wrycza, E. Dahlman, S. Parkvall, J. Skold, HSPA Evolution: The Fundamentals for Mobile Broadband, Academic Press, Oxford, UK, 2014.

[3] H. Holma, A. Toskala, WCDMA for UMTS: HSP Evolution and LTE, John Wiley & Sons, Chichester, UK, 2010.

[4] E. Dahlman, S. Parkvall, J. Skold, P. Beming, 3G Evolution: HSPA and LTE for Mobile Broadband, Academic Press, Oxford, UK, 2010.

[5] S. Sesia, I. Toufik, M. Baker, LTE: The UMTS Long term Evolution, Wiley Online Library, Chichester, UK, 2009.

[6] X. Chu, D. Lopez-Perez, Y. Yang, F. Gunnarsson, Heterogeneous Cellular Networks: Theory, Simulation and Deployment, Cambridge University Press, Cambridge, UK, 2013.

[7] J.T.J. Penttinen, The Telecommunications Handbook: Engineering Guidelines for Fixed, Mobile and Satellite Systems, John Wiley & Sons, Chichester, UK, 2015.

[8] J.G. Andrews, H. Claussen, M. Dohler, S. Rangan, M.C. Reed, Femtocells: past, present, and future, IEEE J. Sel. Areas Commun. 30 (3) (2012) 497–508.

[9] H. Claussen, L.T.W. Ho, L.G. Samuel, An overview of the femtocell concept, Bell Labs Tech. J. 13 (1) (2008) 221–245.

[10] V. Chandrasekhar, J.G. Andrews, A. Gatherer, Femtocell networks: a survey, Commun. Mag. IEEE 46 (9) (2008) 59–67.

[11] A. Kumar, P. Jarich, 3G wireless investment motivations: why do mobile operators continue to invest in 3G wireless network infrastructure? Technical Paper available at http://www.currentanalysis.com/ (September 2015).

[12] Ericsson Company, Ericsson Mobility Report. On The Pulse of The Networked Society. Technical Paper available at http://www.ericsson.com/res/docs/2015/ericsson-mobility-report-june-2015.pdf (June 2015).

[13] J.P. Romero, O. Sallent, R. Agusti, M.A. Diaz-Guerra, Radio Resource Management Strategies in UMTS, John Wiley & Sons, 2005.

[14] M. Stasiak, M. Glabowski, A. Wisniewski, P. Zwierzykowski, Modelling and Dimensioning of Mobile Wireless Networks: From GSM to LTE, John Wiley & Sons, Chichester, UK, 2010.

[15] M. Rahnema, UMTS Network Planning, Optimization, and Inter-operation With GSM, John Wiley & Sons, Chichester, UK, 2008.

[16] S.G. Glisic, Adaptive WCDMA: Theory and Practice, John Wiley & Sons, Chichester, UK, 2003.

[17] C. Chevallier, C. Brunner, A. Garavaglia, K.P. Murray, K.R. Baker, WCDMA (UMTS) Deployment Handbook: Planning and Optimization Aspects, John Wiley & Sons, Chichester, UK, 2006.

[18] R. Ferrus, J. Olmos, H. Galeana, Evaluation of a cell selection framework for radio access networks considering backhaul resource limitations, in: IEEE 18th International Symposium on Personal, Indoor and Mobile Radio Communications, PIMRC 2007, IEEE, 2007, pp. 1–5.

[19] J.J. Olmos, R. Ferrus, H. Galeana-Zapién, Analytical modeling and performance evaluation of cell selection algorithms for mobile networks with backhaul capacity constraints, IEEE Trans. Wirel. Commun. 12 (12) (2013) 6011–6023.

[20] S. Chia, M. Gasparroni, P. Brick, The next challenge for cellular networks: backhaul, Microw. Mag. IEEE 10 (5) (2009) 54–66.

[21] L. Cuadra, S. Salcedo-Sanz, A.D. Carnicer, M.A. Del Arco, J.A. Portilla-Figueras, A novel grouping genetic algorithm for assigning resources to users in WCDMA networks, in: Applications of Evolutionary Computation, Springer, 2015, pp. 42–53.

[22] S. Salcedo-Sanz, A survey of repair methods used as constraint handling techniques in evolutionary algorithms, Comput. Sci. Rev. 3 (3) (2009) 175–192.

[23] F. Neri, C. Cotta, Memetic algorithms and memetic computing optimization: a literature review, Swarm Evolut. Comput. 2 (2012) 1–14.

[24] J.A. Joines, M.G. Kay, Utilizing hybrid genetic algorithms, in: Evolutionary Optimization, Springer, Berlin, Germany, 2002, pp. 199–228.

[25] C. Zhang, J. Chen, B. Xin, Distributed memetic differential evolution with the synergy of Lamarckian and Baldwinian learning, Appl. Soft Comput. 13 (5) (2013) 2947–2959.

[26] Y. Qi, F. Liu, M. Liu, M. Gong, L. Jiao, Multi-objective immune algorithm with Baldwinian learning, Appl. Soft Comput. 12 (8) (2012) 2654–2674.

[27] H. Galeana-Zapién, R. Ferrús, Design and evaluation of a backhaul-aware base station assignment algorithm for OFDMA-based cellular networks, IEEE Trans. Wirel. Commun. 9 (10) (2010) 3226–3237.

[28] E. Falkenauer, The grouping genetic algorithms-widening the scope of the gas, Belg. J. Oper. Res. Stat. Comput. Sci. 33 (1) (1992) 2.

[29] E. Falkenauer, Genetic Algorithms and Grouping Problems, John Wiley & Sons, Inc., Chichester, UK, 1998.

[30] P. De Lit, E. Falkenauer, A. Delchambre, Grouping genetic algorithms: an efficient method to solve the cell formation problem, Math. Comput. Simul. 51 (3) (2000) 257–271.

[31] E.C. Brown, M. Vroblefski, A grouping genetic algorithm for the microcell sectorization problem, Eng. Appl. Artif. Intell. 17 (6) (2004) 589–598.

[32] T. James, M. Vroblefski, Q. Nottingham, A hybrid grouping genetic algorithm for the registration area planning problem, Comput. Commun. 30 (10) (2007) 2180–2190.

[33] L.E. Agustí n-Blas, S. Salcedo-Sanz, P. Vidales, G. Urueta, J.A. Portilla-Figueras, Near optimal citywide WiFi network deployment using a hybrid grouping genetic algorithm, Expert Syst. Appl. 38 (8) (2011) 9543–9556.

[34] C.K. Tan, T.C. Chuah, S.W. Tan, M.L. Sim, Efficient clustering scheme for OFDMA-based multicast wireless systems using grouping genetic algorithm, Electron. Lett. 48 (3) (2012) 184–186.

[35] A. Ganti, T.E. Klein, M. Haner, Base station assignment and power control algorithms for data users in a wireless multiaccess framework, IEEE Trans. Wirel. Commun. 5 (9) (2006) 2493–2503.

[36] M. Dosaranian-Moghadam, H. Bakhshi, G. Dadashzadeh, M. Godarzvand-Chegini, Joint base station assignment, power control error, and adaptive beamforming for DS-CDMA cellular systems in multipath fading channels, in: Mobile Congress (GMC), 2010 Global, IEEE, 2010, pp. 1–7.

[37] G. Dartmann, W. Afzal, X. Gong, G. Ascheid, Joint optimization of beamforming, user scheduling, and multiple base station assignment in a multicell network, in: Wireless Communications and Networking Conference (WCNC), 2011 IEEE, IEEE, 2011, pp. 209–214.

[38] M. Sanjabi, M. Razaviyayn, Z.-Q. Luo, Optimal joint base station assignment and beamforming for heterogeneous networks, IEEE Trans. Signal Process. 62 (8) (2014) 1950–1961.

[39] C. Skianis, Introducing automated procedures in 3G network planning and optimization, J. Syst. Softw. 86 (6) (2013) 1596–1602.

[40] J.A. Portilla-Figueras, S. Salcedo-Sanz, A. Oropesa-Garcí a, C. Bouso no-Calzón, Cell size determination in WCDMA systems using an evolutionary programming approach, Comput. Oper. Res. 35 (12) (2008) 3758–3768.

[41] D. Tsilimantos, D. Kaklamani, G. Tsoulos, Particle swarm optimization for UMTS WCDMA network planning, in: 2008 3rd International Symposium on Wireless Pervasive Computing, ISWPC, IEEE, 2008, pp. 283–287.

[42] J. Nasreddine, J. Pérez-Romero, O. Sallent, Advanced spectrum management for the downlink of WCDMA systems using genetic algorithms, in: Vehicular Technology Conference, 2009. VTC Spring 2009. IEEE 69th, IEEE, 2009, pp. 1–6.

[43] M. Karakoc, A. Kavak, Genetic approach for dynamic OVSF code allocation in 3G wireless networks, Appl. Soft Comput. 9 (1) (2009) 348–361.

[44] F. Gu, H. Liu, M. Li, Evolutionary algorithm for the radio planning and coverage optimization of 3G cellular networks, in: International Conference on Computational Intelligence and Security, CIS'09, 2009, vol. 2, IEEE, 2009, pp. 109–113.

[45] C.-S. Wang, Y.-D. Chen, Base station deployment with capacity and coverage in WCDMA systems using genetic algorithm at different height, in: 2012 Sixth International Conference on Genetic and Evolutionary Computing (ICGEC), IEEE, 2012, pp. 546–549.

[46] L. Shaobo, P. Weijie, Y. Guanci, C. Linna, Optimization of 3G wireless network using genetic programming, in: Second International Symposium on Computational Intelligence and Design, 2009. ISCID'09, vol. 2, IEEE, 2009, pp. 131–134.

[47] M.E. Aydin, R. Kwan, C. Leung, C. Maple, J. Zhang, A hybrid swarm intelligence algorithm for multiuser scheduling in HSDPA, Appl. Soft Comput. 13 (5) (2013) 2990–2996.

[48] S.-f. Zhu, F. Liu, Y.-t. Qi, Z.-y. Chai, J.-s. Wu, Immune optimization algorithm for solving joint call admission control problem in next-generation wireless network, Eng. Appl. Artif. Intell. 25 (7) (2012) 1395–1402.

[49] H.-L. Liu, F. Gu, Y.-m. Cheung, S. Xie, J. Zhang, On solving WCDMA network planning using iterative power control scheme and evolutionary multiobjective algorithm [application notes], Comput. Intell. Mag. IEEE 9 (1) (2014) 44–52.

[50] F. Adachi, M. Sawahashi, K. Okawa, Tree-structured generation of orthogonal spreading codes with different lengths for forward link of DS-CDMA mobile radio, Electron. Lett. 33 (1) (1997) 27–28.

[51] Y.-C. Tseng, C.-M. Chao, Code placement and replacement strategies for wideband CDMA OVSF code tree management, IEEE Trans. Mobile Comput. 1 (4) (2002) 293–302.

[52] S. Kasapović, N. Sarajlić, OVSF code assignment in UMTS networks, in: Microwave and Telecommunication Technology (CriMiCo), 2010 20th International Crimean Conference, IEEE, 2010, pp. 429–432.

[53] Nokia-Siemens Networks. Dimensioning WCDMA RAN. Technical Paper available at http://www.slideshare.net/amini110/dimensioning-wcdma-ran-36325000 (June 2007). RNC3267-trial. Nokia WCDMA RAN, Rel. RA506, System Library, v.1. DN70118376, Issue 2-0 en.

[54] J. Lempiä inen, M. Manninen, UMTS Radio Network Planning, Optimization and QoS Management, Kluwer Academic Publishers, Dodrecht, 2003.

[55] T.K. Sarkar, Z. Ji, K. Kim, A. Medouri, M. Salazar-Palma, A survey of various propagation models for mobile communication, Antennas and Propag. Mag. IEEE 45 (3) (2003) 51–82.

[56] T.S. Rappaport, et al., Wireless Communications: Principles and Practice, vol. 2, Prentice Hall PTR, New Jersey, 1996.

[57] A. Tahat, Y. Alqudah, et al., Analysis of propagation models at 2.1 GHz for simulation of a live 3G cellular network, in: Wireless Advanced (WiAd), 2011, IEEE, 2011, pp. 164–169.

[58] N. Belhadj, B. Oueslati, T. Aguili, Adjustment of Cost231 Walfisch-Ikegami model for HSPA+ in Tunisian urban environments, in: 2015 2nd World Symposium on Web Applications and Networking (WSWAN), IEEE, 2015, pp. 1–6.

[59] T. Acar, F. Caliskan, E. Aydin, Comparison of computer-based propagation models with experimental data collected in an urban area at 1800 MHz, in: Wireless and Microwave Technology Conference (WAMICON), 2015 IEEE 16th Annual, IEEE, 2015, pp. 1–6.

[60] K. Uchida, N. Hadano, M. Takematsu, J. Honda, Propagation estimation by using building coverage and floor area ratios based on 1-ray model combined with Okumura-Hata, in: 2014 17th International Conference on Network-Based Information Systems (NBiS), IEEE, 2014, pp. 555–560.

[61] A. Aragon-Zavala, Antennas and Propagation for Wireless Communication Systems, John Wiley & Sons, Chichester, UK, 2008.

[62] C.A. Coello, Theoretical and numerical constraint-handling techniques used with evolutionary algorithms: a survey of the state of the art, Comput. Methods Appl. Mech. Eng. 191 (11) (2002) 1245–1287.

[63] G. Gréwal, S. Coros, D.K. Banerji, A. Morton, Comparing a genetic algorithm penalty function and repair heuristic in the DSP application domain, Artif. Intell. Appl. 3 (2006) 1–39.

[64] T. Bäck, M. Schütz, S. Khuri, A comparative study of a penalty function, a repair heuristic, and stochastic operators with the set-covering problem, in: Artificial Evolution, Springer, 1996, pp. 320–332.

[65] T. Runnarson, X. Yao, Constrained evolutionary optimization-the penalty function approach, Evolut. Optim. 8 (2002) 7–113.

[66] D. Orvosh, L. Davis, Using a genetic algorithm to optimize problems with feasibility constraints, in: Proceedings of the First IEEE Conference on Evolutionary Computation, 1994. IEEE World Congress on Computational Intelligence, IEEE, 1994, pp. 548–553.

[67] D. Orvosh, L. Davis, Shall we repair? genetic algorithms combinatorial optimization and feasibility constraints, in: Proceedings of the 5th International Conference on Genetic Algorithms, Morgan Kaufmann Publishers Inc., 1993, p. 650.

[68] P. Chootinan, A. Chen, Constraint handling in genetic algorithms using a gradient-based repair method, Comput. Oper. Res. 33 (8) (2006) 2263–2281.

[69] G.G. Mitchell, D. O?Donoghue, D. Barnes, M. McCarville, GeneRepair – a repair operator for genetic algorithms, in: GECCO, Citeseer, 2003, pp. 235–239.

[70] J. Gottlieb, G.R. Raidl, The effects of locality on the dynamics of decoder-based evolutionary search, in: GECCO, 2000, pp. 283–290.

[71] D.N. Skoutas, A.N. Rouskas, A Scheduling algorithm with dynamic priority assignment for WCDMA systems, IEEE Trans. Mobile Comput. 8 (1) (2009) 126–138.

[72] J. Laiho, A. Wacker, T. Novosad, Radio Network Planning and Optimisation for UMTS, John Wiley & Sons, Chichester, UK, 2006.

Available online at www.sciencedirect.com

ScienceDirect

Solar Energy 132 (2016) 129–142

SOLAR ENERGY

www.elsevier.com/locate/solener

A novel Grouping Genetic Algorithm–Extreme Learning Machine approach for global solar radiation prediction from numerical weather models inputs

A. Aybar-Ruiz[a], S. Jiménez-Fernández[a], L. Cornejo-Bueno[a], C. Casanova-Mateo[b], J. Sanz-Justo[b], P. Salvador-González[b], S. Salcedo-Sanz[a,*]

[a] *Department of Signal Processing and Communications, Universidad de Alcalá, Alcalá de Henares, Spain*
[b] *LATUV: Remote Sensing Laboratory, Universidad de Valladolid, Valladolid, Spain*

Received 17 July 2015; received in revised form 24 February 2016; accepted 3 March 2016

Communicated by: Associate Editor Jan Kleissl

Abstract

This paper presents a novel scheme for global solar radiation prediction, based on a hybrid neural-genetic algorithm. Specifically a grouping genetic algorithm (GGA) and an Extreme Learning Machine algorithm (ELM) have been merged in a single algorithm, in such a way that the GGA solves the optimal selection of features, and the ELM carries out the prediction. The proposed scheme is also novel because it uses as input of the system the output of a numerical weather meso-scale model (WRF), i.e., atmospherical variables predicted by the WRF at different nodes. We consider then different problems associated with this general algorithmic framework: first, we evaluate the capacity of the GGA–ELM for carrying out a statistical downscaling of the WRF to a given point of interest (where a measure of solar radiation is available), i.e., we only take into account predictive variables from the WRF and the objective variable at the same time tag. In a second evaluation approach, we try to predict the solar radiation at the point of interest at different time tags $t+x$, using predictive variables from the WRF. Finally, we tackle the complete prediction problem by including previous values of measured solar radiation in the prediction. The proposed algorithm and its efficiency for selecting the best set of features from the WRF are analyzed in this paper, and we also describe different operators and dynamics for the GGA. Finally, we evaluate the performance of the system with these different characteristics in a real problem of solar radiation prediction at Toledo's radiometric observatory (Spain), where the proposed system has shown an excellent performance in all the subproblems considered, in terms of different error metrics.

Keywords: Grouping genetic algorithm (GGA); Global solar radiation prediction; Solar energy; Extreme Learning Machines

1. Introduction

Solar energy is an important source of renewable and clean energy, currently under expansion in different countries of the world, and with a huge potential to contribute significantly to the energy mix and nations' economies of these countries. Solar energy development is specially important in mid-east and southern Europe countries, where the solar resource can be better exploited all year around (Kalogirou, 2014). Solar production is intrinsically stochastic (with reference to intra-hour solar forecasting) and significant variations in solar energy production occur due to the presence of clouds, atmospheric dust or

* Corresponding author at: Department of Signal Processing and Communications, Universidad de Alcalá, 28871 Alcalá de Henares, Madrid, Spain. Tel.: +34 91 885 6731; fax: +34 91 885 6699.
E-mail address: sancho.salcedo@uah.es (S. Salcedo-Sanz).

http://dx.doi.org/10.1016/j.solener.2016.03.015
0038-092X/© 2016 Elsevier Ltd. All rights reserved.

particles. For longer time horizons (e.g. 6 h or more), physics-based models are usually employed (Diagne et al., 2014; Coimbra et al., 2013; Inman et al., 2013). Because of this, prediction of the energy production in solar energy plants is an important problem to integrate this renewable energy in the system. The problem of solar energy prediction usually involves the accurate prediction of the solar radiation at a given point (the solar plant facility), and this prediction depends completely on different atmospheric variables (Inman et al., 2013; Khatib et al., 2012; Sozen et al., 2004; Voyant et al., 2011).

In the last years, many different approaches have been proposed for global solar radiation prediction, a lot of them using Machine Learning or Computational Intelligence techniques. The majority of these approaches include different inputs based on geographical and atmospheric parameters such as latitude, longitude, temperature, wind speed and direction, sunshine duration, and precipitation (Mellit and Kalogirou, 2008; Mubiru, 2008). According to Bilgili and Ozoren (2011), sunshine duration, air temperature and relative humidity are the most widely used meteorological parameters to predict daily solar radiation and its components. All these parameters are well correlated with the daily solar global radiation as pointed out in Yacef et al. (2012). In López et al. (2005) a Bayesian framework for artificial neural networks, named as automatic relevance determination method, was developed to evaluate the more relevant input parameters in modelling solar radiation. In fact, neural computation paradigm has been massively applied to this prediction problem, like in Benghanem and Mellit (2010), where it is shown that Radial Basis Functions (RBF) neural networks obtain excellent performance in the estimation of solar radiation. In Dorvlo et al. (2002) a comparison between Multi-Layer Perceptrons (MLP) and RBF neural networks in a problem of solar radiation estimation is carried out. Experiments in eight stations in Oman show the good results obtained with the neural algorithms. A similar approach, also comparing MLPs and RBFs (with different predictive variables) has been recently proposed in Behrang et al. (2010), in this case the authors test the neural network with data obtained in Iran. In Paoli et al. (2010) the performance of a MLP in a problem of solar radiation prediction in time series is compared to that of ARIMA, Bayesian inference, Markov Chains and k-Nearest Neighbors models, for specific problems in Corsica and Southern France. Another work dealing with solar radiation time series prediction is Wu and Chan (2011), where a hybrid algorithm that involves an ARMA model and a time-delay neural network is proposed. In Hocaoglu et al. (2008) a neural network to predict hourly solar radiation in a region of Turkey is proposed. The paper also introduces a 2D model for solar radiation useful for visualization and data inspection. In Fu and Cheng (2013), the forecasting of solar irradiance proposed utilizes features extracted from all-sky images, such as the number of cloud pixels, frame difference, gradient magnitude, intensity level, accumulated intensity along the vertical line of sun or the number of corners in the image. Other works on solar radiation prediction involve ARMA models, as Ji and Chee (2011) a hybrid approach based on ARMA and time delay neural networks has been successfully tested in data from a solar station in Singapore. Another paper involving hybrid ARMA and neural networks is (Voyant et al., 2013), where this hybrid approach is successfully applied to solar radiation prediction in different cities of the French Mediterranean coast and Corsica. Alternative approaches that apply neural networks as prediction methodology also include novel predictive variables, such as satellite data (Senkal and Kuleli, 2009) or temperature and relative humidity (Rehman and Mohandes, 2008). Other machine learning algorithms, such as Support Vector Regression (SVR) algorithms have been also applied to solar radiation prediction problems from meteorological predictive variables (Chen et al., 2011; Zeng and Qiao, 2013). Specifically, a least-square SVM is proposed in that work, comparing the results obtained with that of auto-regressive and RBF neural networks. In Rahimikoob (2010) the potential of multi-layer perceptron neural networks with back-propagation training algorithm is shown in a problem of global solar radiation estimation in Iran. Results comparing the performance of the neural networks with that of an empirical equation for global solar radiation prediction (Hargreaves and Samani equation) show good performance of the neural approach. In Bhardwaj et al. (2013) a hybrid approach that includes hidden Markov models and generalized fuzzy models has been proposed and tested in real solar irradiation data in India. Finally, we discuss very recent hybrid approaches proposed to problems of solar energy prediction, such as Olatomiwa et al. (2015) where a SVR has been hybridized with a firefly algorithm to select the best parameters of the SVR, or Mohammadi et al. (2015), where a hybrid SVR-Wavelets approach is presented in a problem of horizontal global solar radiation prediction. The goodness of this novel approach has been tested in a real problem of solar radiation estimation in Bandar Abbas (Iran). Moreover, in Diagne et al. (2014), a post-processing technique (Kalman filtering) is used to improve the hour-ahead forecasted Global Horizontal Irradiance (GHI) from (1) the measured GHI at the ground, and (2) the Weather Research and Forecasting (WRF) meso-scale model, and results at Reunion Island are provided.

Different approaches discussing Extreme Learning Machine (ELM, a novel training method for artificial neural networks) applications in solar radiation prediction problems have been recently proposed, such as Sahin et al. (2014), where the ELM approach is applied to a solar radiation prediction problem from satellite measures. In Alharbi (2013) a case study of solar radiation prediction in Saudi Arabia is discussed comparing the performance of artificial neural networks with classical training and ELMs. In Dong et al. (2014) a hybrid wavelet-ELM approach is tested in a problem of solar radiation prediction for application in a photovoltaic power station.

Finally, in Salcedo-Sanz et al. (2013) a comparison of a support vector regression algorithm and an ELM is carried out in a problem of direct solar radiation prediction, with application in solar thermal energy systems, and in Salcedo-Sanz et al. (2014), where a hybrid ELM-Coral Reefs Optimization is proposed for solar radiation prediction in Southern Spain.

In spite of this impressive amount of previous works on Machine Learning approaches for solar radiation prediction problems, there is still margin for improvement. Note that very few papers among those discussed above use numerical weather models outputs as inputs for prediction, and this strategy has been successfully applied before in wind speed prediction (Salcedo-Sanz et al., 2009,). Thus, the use of numerical weather models' prediction to feed Machine Learning approaches in solar energy prediction is a promising field that can produce efficient prediction approaches. Moreover, the use of different meta-heuristics to perform feature selection in different prediction problems has been successfully reported in the literature, and (Salcedo-Sanz et al., 2002, 2014) refer the use of evolutionary-type meta-heuristics. Nevertheless, the use of novel evolutionary paradigms such as grouping genetic algorithms (GGAs) has not been tested before, to our knowledge, for feature selection in solar energy applications, thus representing a novelty of this work. GGA is useful for feature selection since it is able to group different sets of features and evaluate these sets under different objective functions (best, average or even worst error measures can be used to guide the genetic search). Also, the use of specific GGA operators is another novel characteristic that has not been evaluated before in previous works dealing with feature selection problems.

The objective of this paper is twofold: first, we consider a problem of global solar radiation prediction from numerical weather models, specifically the WRF meso-scale model. WRF provides a prediction of atmospherical variables at different pressure levels in a given zone, that will be used as inputs in a prediction system to estimate the global solar radiation at a different point. The second contribution of the paper is the development of a hybrid grouping genetic algorithm–ELM algorithm to carry out this global solar radiation prediction. The grouping genetic algorithm (GGA) proposed will perform a process of feature selection, focused on filtering the best features from the WRF model to do the prediction, whereas the ELM approach will do the final prediction of the global solar radiation at a given point, using the features selected by the GGA. We will discuss in detail the proposed algorithm, giving some variants of its dynamics, that lead to different performances. With this algorithmic framework in mind, we then tackle a number of subproblems related to global radiation prediction: the first problem considered consists of predicting the solar radiation registered in a given point $\mathcal{P}$ at time $t + x$ (for $x = 0, \ldots, 3$), using as predictive variables the set $\mathcal{V}$, or any subset of it. Note that when $x = 0$, the problem is known as *statistically downscaling* the solar radiation prediction of model $\mathcal{M}$ to point $\mathcal{P}$. The ultimate goal of this approach for $x = 0$ (SubProblem 1) is to evaluate what features (predictive variables) from the Numerical Model are useful for this prediction. Note that for $x > 0$ (SubProblem 2) we are evaluating the prediction performance of the system only using as predictive variables the outputs of the WRF. Finally, we tackle in this paper a forecasting problem that takes into account data from the Numerical Model and also objective variable data measured at the measuring station considered (SubProblem 3). This last subproblem uses the best predictive variables set found in SubProblem 1. Results for all these subproblems using real data from Toledo's radiometric station (Spain), will be discussed in the experimental part of the paper.

The structure of the rest of the paper is the following: next section presents the problem formulation and describes the WRF meso-scale model input variables involved. Section 3 describes the proposed hybrid GGA–ELM approach. Section 4 presents the main results obtained in a real solar radiation prediction problem at Toledo, Spain. Finally, Section 5 closes the paper by giving some concluding remarks.

2. Problem formulation

The problem considered in this paper can be stated in the following way: let $\mathcal{P}$ be a given location of the Earth's surface where the global solar radiation ($\mathcal{I}_t$) must be predicted ($\hat{\mathcal{I}}_t$), at a given time t. To do this, let us consider the output, $\mathcal{V}$, of a numerical meso-scale model $\mathcal{M}$, in a number M of nodes, consisting of the prediction at time t for N atmospheric variables, $\mathcal{V} = (\varphi_{11}, \ldots, \varphi_{1N}, \varphi_{21}, \ldots, \varphi_{2N}, \ldots, \varphi_{M1}, \ldots, \varphi_{MN})$, as shown in Fig. 1.

Note that $\mathcal{M}$ may provide an atmospheric variable at the ground level, or at the ground level and also at different pressure levels. In the latter, each pressure level is considered as a different variable n at node m, φ_{mn}.

SubProblem 1 deals with the prediction of the global solar radiation registered in $\mathcal{P}$ at time t, using as predictive variables the set $\mathcal{V}$, or any subset of it. This type of problems is usually known in other works as *statistically downscaling* the solar radiation prediction of model $\mathcal{M}$ to point $\mathcal{P}$.

SubProblem 2 increases the forecast horizon, predicting the global solar radiation in $\mathcal{P}$ at time $t + x$ (for

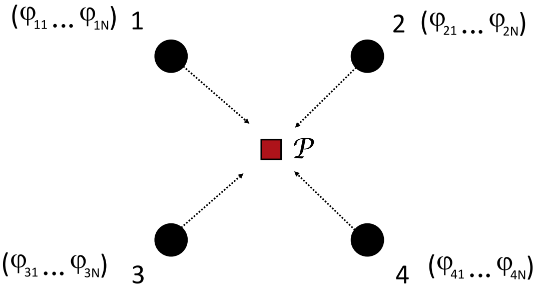


Fig. 1. Solar radiation prediction scheme used in this work for $M = 4$.

$x = 1, 2, \ldots, \mathcal{X}$), considering the set $\mathcal{V}$ (or any subset) as predictive variables.

Finally, SubProblem 3 analyzes a forecasting problem at $\mathcal{P}$ considering previous radiation values. For this purpose, the best set of features found by the GGA–ELM in SubProblem 1, i.e. $\mathcal{V}^* = (\varphi_1^*, \ldots, \varphi_{\mathcal{K}}^*)$, where $\mathcal{K}$ is the number of optimal features obtained by the GGA–ELM approach, have been used. In addition, we consider objective variable data measured at point $\mathcal{P}$ for previous time tags ($\mathcal{I}_{t-z}$), for $z = 1, \ldots, \mathcal{Z}$, in the process of $\hat{\mathcal{I}}_{t+x}$ forecasting, for $x = 1, 2, \ldots, \mathcal{X}$.

2.1. Location under study and objective variable data

This work predicts the global solar radiation at a given location $\mathcal{P}$ that pinpoints a Meteorological State Agency of Spain (AEMET) station sited in Toledo (39° 53′ 5″N, 4° 02′ 43″W). Toledo's measuring station is located in the South Plateau of the Iberian Peninsula (See Fig. 2), around 75 km south of Madrid (the capital of Spain) at an altitude of 515 m.

According to AEMET's Climate Summary Guide (1981–2010), Toledo has an annual mean temperature of

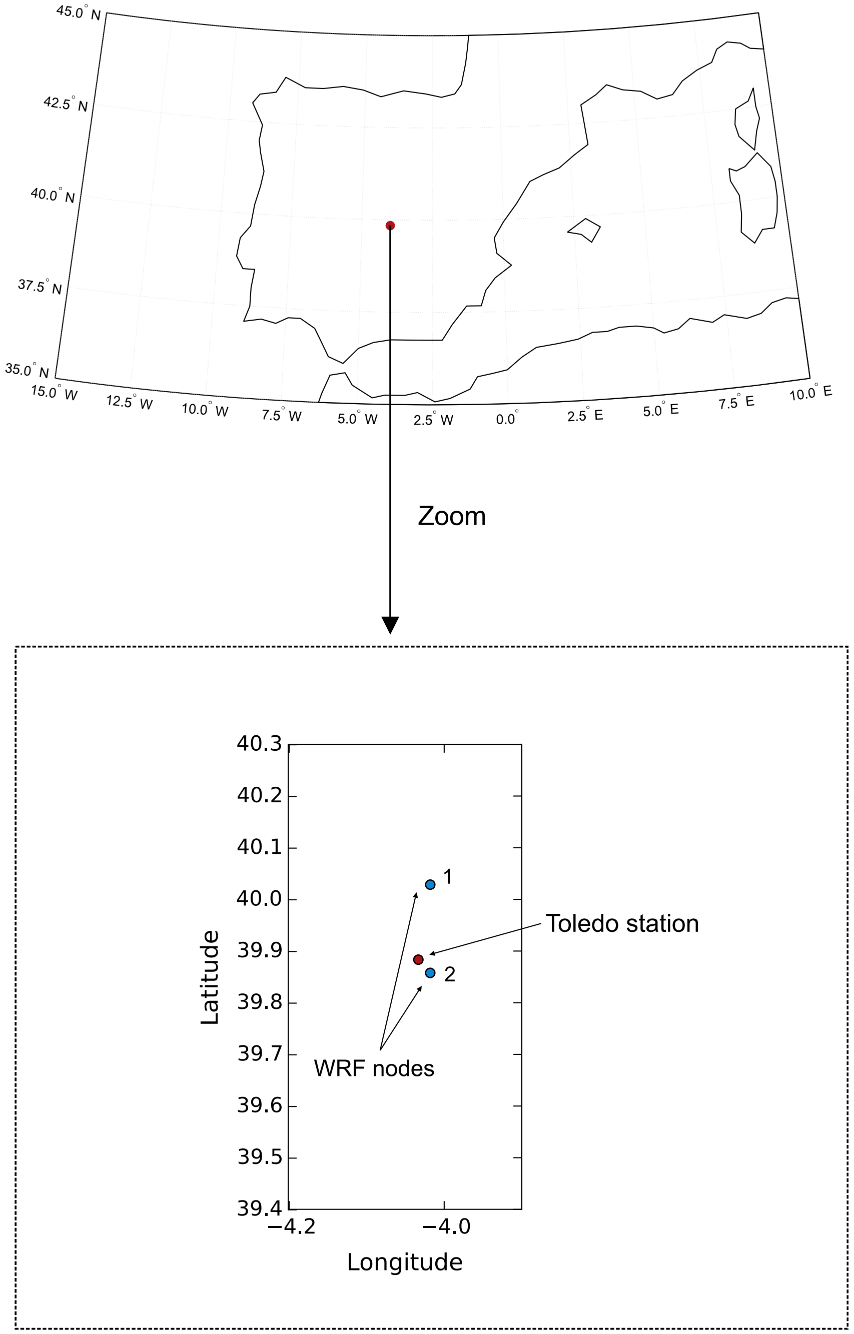


Fig. 2. Location of the Toledo's measuring station and the $M = 2$ WRF nodes considered for the prediction.

15.8 °C, dry summers, an annual mean precipitation of 342.2 mm, and an annual mean water vapor tension of 10.8 hPa. In the light of these figures and according to the Köppen climate classification, Toledo could be classified as a Csa climate (Interior Mediterranean: Mild with dry, hot summer). Regarding cloud cover, 27.7% annual days are categorized as cloud free, 53.2% present sky conditions categorized as few or scattered clouds, and the remaining 19.1% are categorized as broken or overcast.

As objective variable data to train and test the algorithms, we consider one year of hourly global solar radiation data (from May 1st, 2013 to April 30th, 2014) collected at Toledo's measuring station. To measure global solar radiation, a Kipp & Zonen CMP11 Pyranometer is used. All radiation measurements gathered by the AEMET are made following the World Meteorological Organization (WMO) standards included in the WMO Guide to Meteorological Instruments and Methods of Observation (2008 edition, updated in 2010).

2.2. Model $\mathcal{M}$: the Weather Research and Forecasting model (WRF)

In this paper we use the well-known Weather Research and Forecasting (WRF) meso-scale model as $\mathcal{M}$ (Skamarock et al., 2005). WRF is an extremely powerful meso-scale numerical weather prediction system designed for atmospheric research and also for operational forecasting needs. The WRF was developed in collaboration by the National Center for Atmospheric Research (NCAR), the National Centers for Environmental Prediction (NCEP), the Forecast Systems Laboratory (FSL), the Air Force Weather Agency (AFWA), the Naval Research Laboratory, the University of Oklahoma, and the Federal Aviation Administration (FAA) of the USA. The WRF has been used in a wide range of meteorological (Giannaros et al., 2015) and renewable energy applications (Carvalho et al., 2014).

In our study, WRF model version 3.6 has been used. It has been run every 12 h since it was started in 2011. Meteorological data are calculated over a window ranging in latitude from 34° 33′ 43″N to 44° 28′ 12″N, and in longitude from 4° 25′ 12″W to 4° 23′ 2″E. In this window, the grid has 99 elements from West to East, and 59 elements from North to South, roughly, each grid element covers 15×30 km^2. Atmospheric values are calculated, in the vertical dimension, at 37 levels above the ground, at ground level, and at four additional levels beneath the surface. The grid type is Arakawa, that is to say that data are calculated at the center of each element, with a 72 s time step.

WRF is initialized by data coming from NCEP FNL Operational Global Analysis and works in non-hydrostatic way. The short wave scheme used is that from MM5 shortwave (Dudhia), and the long wave model is the RRTM (Rapid Radiative Transfer Model). A radiation time step of 30 min was applied to each radiation domain. The land surface fluxes were obtained by Monin–Obukhov similarity theory, the surface physics was solved by the Unified Noah land surface model and the Planetary Boundary Level (PBL) by means of the Yonsei University (YSU) PBL scheme. The PBL was calculated at every basic time step, and five layers were considered in land surface model. Cumulus retrieval parameters was done by using the new Kain–Fritsch scheme, as in MM5 and Eta/NMM ensemble version, with a time step of 5 min.

Finally, microphysics was carried out by the WSM 3-class scheme and the turbulent diffusion option was to select 2nd order diffusion on model levels. This complements vertical diffusion done by the PBL scheme.

WRF output at two points located at (39° 51′N, 4° 01′W) and (40° 01′N, 4° 01′W), are selected as predictive variables (see Fig. 2). Specifically, Table 1 shows the 46 variables considered for each of these points, summing up a total of 92 predictive variables.

The main variables considered are the following (at different pressure levels):

- OLR: the top of atmosphere outgoing long-wave radiation (W/m^2).
- GLW: the downward long-wave flux at ground surface (W/m^2).
- SWDOWN: the downward short-wave flux at ground surface (W/m^2).
- u: the horizontal wind component in the x direction at different pressure levels (m/s).
- v: the horizontal wind component in the y direction different pressure levels (m/s).
- CLDFRA: the fraction of clouds in each cell. Cloud fraction ranges from 0 (no clouds) to 1 (clouds in a spatial grid cell).
- QVAPOR: the water vapor mixing ratio (in kg/kg). This variable is defined as the ratio of the mass of a water vapor to the mass of dry air.
- T: the temperature (in K) at different levels. Note that this variable is not directly provided by the WRF model. Thus, we have obtained it from the WRF perturbation potential temperature (T') which, in turn, is related with the potential temperature (θ) through the relation $\theta = T' + 300$. Potential temperature is simply defined as the temperature that an unsaturated parcel of dry

Table 1
Predictive variables used in the experiments (46 variables per node of the WRF model).

Variable	Units	Pressure levels (hPa)
OLR	W/m^2	–
GLW	W/m^2	Ground
SWDOWN	W/m^2	Ground
u	m/s	Ground, 850, 700, 500, 400, 300, 200, 100, 50
v	m/s	Ground, 850, 700, 500, 400, 300, 200, 100, 50
CLDFRA	1/0	Ground, 850, 700, 500, 400, 300, 200
QVAPOR	kg/kg	Ground, 850, 700, 500, 400, 300, 200, 100, 50
Temperature	K	Ground, 850, 700, 500, 400, 300, 200, 100, 50

air would have if brought adiabatically and reversibly from its initial state to a standard pressure, P_0, typically 100,000 Pa. Its mathematical expression is as shown in Eq. (1), where κ is the Poisson constant.

$$T = \theta\left(\frac{P}{P_0}\right)^{\kappa} \tag{1}$$

3. The Grouping Genetic Algorithm

The grouping genetic algorithm (GGA) is a class of evolutionary algorithm especially modified to tackle grouping problems. In these problems, a certain set of predefined groups bring together a certain number of items (i.e. in feature selection problems, several subsets of features). It was first proposed by Falkenauer (1992, 1998), who realized that traditional genetic algorithms had difficulties when they were applied to grouping problems (mainly, the standard binary encoding increases the space search size in this type of problems). GGAs have shown interesting performances on different problems and applications (Agustín-Blas et al., 2011; Brown and Sumichrast, 2005; James et al., 2007a,b). Note that in the GGA, the encoding, crossover and mutation operators of traditional Genetic Algorithms (GA) are modified to obtain a compact algorithm with good performance in grouping problems. In this paper we show how to apply the GGA to solve a feature selection problem in a context of solar radiation prediction. We structure the description of the GGA in Encoding, Operators and Fitness Function calculation.

3.1. Problem encoding

The GGA initially proposed by Falkenauer is a variable-length genetic algorithm. The encoding is carried out by separating each individual in the algorithm into two parts: the first one is an *assignment* part that associates each item to a given group. The second one is a *group* part, that defines which groups must be taken into account for the individual. In problems where the number of groups is not previously defined, it is easy to see why this is a variable-length algorithm: the group part varies from one individual to another, as each individual may contain a different number of groups. In our implementation, an individual $\mathbf{c}$ has the form $\mathbf{c} = [\mathbf{a}|\mathbf{g}]$. Example 1 shows an individual in the proposed GGA presenting 10 features and 4 groups, where group #1 includes features $\{1, 3, 9\}$, group #2 features $\{2, 4, 10\}$, group #3 features $\{5, 7\}$ and, finally, group #4 includes feature $\{6, 8\}$.

Example 1. 1 2 1 2 3 4 3 4 1 2 | 1 2 3 4

In this work, every individual's assignment part is composed of 92 elements, each corresponding to one of the predictive variables provided by the WRF model, as explained in Section 2. Then, the group part is composed of a variable number of elements (groups), resulting in one individual having N_{G_1} groups, while another one in the same population may have N_{G_2} groups.

3.2. Genetic operators

In this paper, the selection operator used is a tournament-based mechanism, similar to the one described in Yao et al. (1999), as it has been shown to be one of the most effective selection operators, avoiding super-individuals and performing an excellent exploration of the search space.

Regarding the crossover operator, different versions of this operator have been implemented and the results obtained compared. For a clearer description, notation used refers to those individuals acting as parents as P_i $(i = 1, 2)$ and those individuals acting as offsprings as O_i $(i = 1, 2)$. The assignment and groups part of a certain individual are referred as A_{P_i} and G_{P_i} (for the parents), or A_{O_i} and G_{O_i} (for the offsprings).

The first crossover operator applied follows the guidelines initially proposed by Falkenauer (1992, 1998), that leads to a two-parents/one-child mechanism. The process (outlined in Fig. 3) carried out by this first crossover operator, $\mathcal{C}_1$, is the following:

1. Randomly choose two parents from the current population: P_1 and P_2. The offspring individual, O_1, is initialized to be equal to P_2.

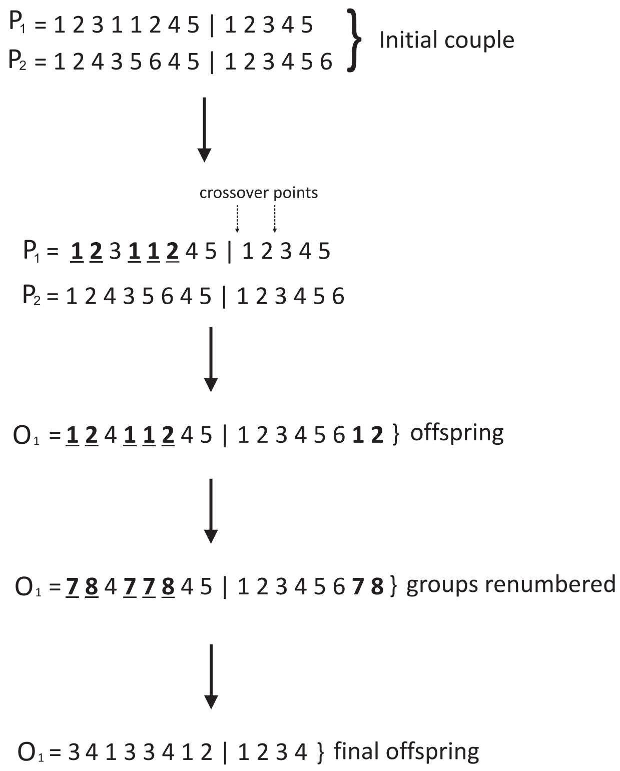

P₁ = 1 2 3 1 1 2 4 5 | 1 2 3 4 5
P₂ = 1 2 4 3 5 6 4 5 | 1 2 3 4 5 6
Initial couple
crossover points
P₁ = 1 2 3 1 1 2 4 5 | 1 2 3 4 5
P₂ = 1 2 4 3 5 6 4 5 | 1 2 3 4 5 6
O₁ = 1 2 4 1 1 2 4 5 | 1 2 3 4 5 6 1 2 } offspring
O₁ = 7 8 4 7 7 8 4 5 | 1 2 3 4 5 6 7 8 } groups renumbered
O₁ = 3 4 1 3 3 4 1 2 | 1 2 3 4 } final offspring

Fig. 3. Outline of the grouping crossover, $\mathcal{C}_1$, implemented in the proposed GGA.

2. Randomly select, for the crossover, two points from G_{P_1}. These two cross-points mark down those groups in-between them, and those features assigned to these groups are selected. In the example presented in Fig. 3, the two crossover points select two groups: group number 1 (G_1) and group number 2 (G_2). Note that, in this case, the features of P_1 belonging to groups G_1 and G_2 are 1, 2, 4, 5, and 6 (marked bold and underlined).
3. Insert those P_1's selected features (in their own positions) in O_1. Then, attach at the end of O_1's group section those new groups inherited from P_1. In the example, it can be seen that the assignment of the features 1, 2, 4, 5 and 6 of O_1 has been inherited from P_1, while the rest of the nodes' assignment has been inherited from P_2.
4. Rename G_{O_1}'s groups to remove duplicates (note that the offspring may have inherited same groups' numbering from both parents). In the example, $G_{O_1} = 1\ 2\ 3\ 4\ 5\ 6\ 1\ 2$ is changed to $G_{O_1} = 1\ 2\ 3\ 4\ 5\ 6\ 7\ 8$. Therefore, A_{O_1} has to be modified accordingly.
5. Remove, empty groups in O_1, if present. In the example considered, it is found that O_1's groups 1, 2, 3, and 6 are empty (there are no features belonging to them), so we can eliminate these groups' identification number and rearrange the rest accordingly. The final offspring is then obtained.

We realized that $\mathcal{C}_1$ crossover operator produces a significant increment of the number of groups after a number of generations of the GGA, so we have tried to correct this point by introducing an alternative grouping crossover. Accordingly, a two-parents/two-children crossover, $\mathcal{C}_2$, is presented. The $\mathcal{C}_2$ process is shown in Fig. 4, and can be described as follows:

1. Randomly choose two parents from the current population: P_1 and P_2.
2. Randomly select, for the crossover, two points from G_{P_1} and two points from G_{P_2}. For each parent, these cross-points mark down those groups in-between them, and those features assigned to these groups are selected.
3. To build G_{O_1}, use the selected section of G_{P_1}. In the example, P_1's selected groups are G_2 and G_3, resulting in the offsprings group part $G_{O_1} = 2\ 3$.To build A_{O_1} use the selected features inherited from P_1.
4. If necessary, rename G_{O_1}'s groups so that groups' numbering starts at 1.
5. Randomly allocate among the offspring's groups those blank features. The final first offspring is then obtained.
6. Repeat steps 2 to 5 using the second parent to obtain the second offspring.

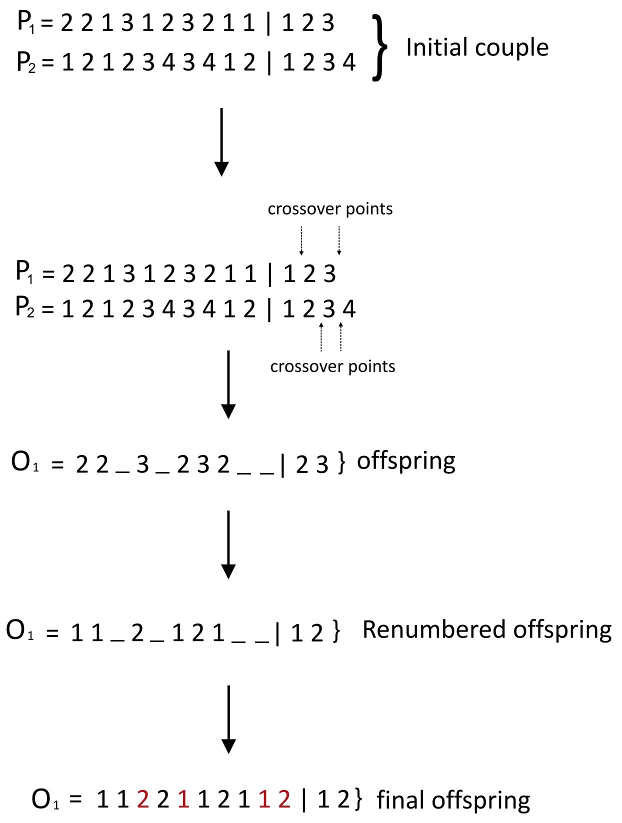

P1 = 2 2 1 3 1 2 3 2 1 1 | 1 2 3
P2 = 1 2 1 2 3 4 3 4 1 2 | 1 2 3 4
Initial couple
crossover points
P1 = 2 2 1 3 1 2 3 2 1 1 | 1 2 3
P2 = 1 2 1 2 3 4 3 4 1 2 | 1 2 3 4
crossover points
O1 = 2 2 _ 3 _ 2 3 2 _ _ | 2 3 } offspring
O1 = 1 1 _ 2 _ 1 2 1 _ _ | 1 2 } Renumbered offspring
O1 = 1 1 2 2 1 1 2 1 1 2 | 1 2} final offspring

Fig. 4. Outline of the grouping crossover, $\mathcal{C}_2$, implemented in the proposed GGA.

Regarding mutation operator, a swapping mutation in which two items are interchanged is applied. Thus, resulting in the assignment of features to different groups. This procedure is carried out with a very low probability ($P_m = 0.01$), to avoid increasing the random search in the process (Falkenauer, 1998).

3.3. Fitness function: the Extreme Learning Machine for feature selection

Two main approaches may be applied for feature selection (Blum and Langley, 1997; Kohavi and John, 1997): wrappers and filters. A wrapper approach uses the direct output of the regressor as objective function, while filter methods apply an external measure (for example, Mutual Information), thus resulting in the algorithm's performance depending completely on the measure selected. Filter methods are usually faster than wrapper methods. On the other hand, wrapper methods have been shown to be more accurate.

In this work we consider wrapper feature selection based on a GGA. Given a specific group of features a prediction of the global solar irradiance must be performed to analyze the goodness of the group of features. For this purpose, the regressor chosen must be as accurate as possible, and also very fast in its training process, in order to avoid high computational burden for the complete algorithm. The *Extreme Learning Machine* (ELM) (Huang et al., 2006, 2011) is a very fast learning method, based on the structure of multi-layer perceptrons, that shows excellent performance in classification and regression problems, and it has been the regressor chosen to be optimized by means of the feature selection process carried out by the GGA.

The most significant property of the ELM training, and one of the reasons why it is so fast, is that it is trained just by randomly setting the network weights, and then obtaining the inverse of the hidden-layer output matrix (Huang et al., 2006, 2012).

Given a training set $\Gamma = \{(\mathbf{x}_i, \mathbf{y}_i) | \mathbf{x}_i \in \mathbb{R}^n, \mathbf{y}_i \in \mathbb{R}^m, i = 1, \ldots, N\}$, where $\mathbf{x}$ stands for the predictive variables, and $\mathbf{y}$ stands for the objective variable. An activation function, $a(x)$, and number of hidden nodes, $\tilde{N}$, the ELM works as follows:

(1) Randomly assign inputs weights $\mathbf{w}_i$ and bias $b_i, i = 1, \ldots, \tilde{N}$.

(2) Calculate the hidden layer output matrix $\mathbf{H}$, defined as

$$\mathbf{H} = \begin{bmatrix} a(\mathbf{w}_1 \cdot \mathbf{x}_1 + b_1) & \cdots & a(\mathbf{w}_{\tilde{N}} \cdot \mathbf{x}_1 + b_{\tilde{N}}) \\ \vdots & \cdots & \vdots \\ a(\mathbf{w}_1 \cdot \mathbf{x}_N + b_1) & \cdots & a(\mathbf{w}_{\tilde{N}} \cdot \mathbf{x}_N + b_{\tilde{N}}) \end{bmatrix}_{N \times \tilde{N}} \tag{2}$$

(3) Compute the output weight vector β as

$$\beta = \mathbf{H}^{\dagger}\mathbf{Y}, \tag{3}$$

where $\mathbf{H}^{\dagger}$ stands for the Moore–Penrose inverse of matrix $\mathbf{H}$ (Huang et al., 2006), and $\mathbf{Y}$ is the training output vector, $\mathbf{Y} = [\mathbf{y}_1, \cdots, \mathbf{y}_N]^T$.

Note that the number of hidden nodes ($\tilde{N}$) is a free parameter of the ELM training, and must be estimated to obtain good results. Usually, scanning a range of $\tilde{N}$ values is the solution for this problem.

The fitness function considered for each element of the GGA is obtained by using the Root Mean Square Error (RMSE) of the prediction. RMSE is used in this work instead of other validation metrics, because large forecast errors and outliers are weighted more strongly than smaller errors, as the latter are more tolerable in solar radiation prediction (Beyer et al., 2011; Kleissl, 2013).

The RMSE formula is shown in Eq. (4), where $\mathcal{I}_t$ stands for the global solar radiation measured at a time t, $\hat{\mathcal{I}}_t$ stands for the global solar radiation estimated by the ELM, and T stands for the number of samples in the test set.

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(\hat{\mathcal{I}}_t - \mathcal{I}_t)^2} \tag{4}$$

Note that RMSE is the only metric used at the ELM's training phase. Nevertheless, to assess the performance of the network (in the test phase), alternative metrics have been also used (see SubSection 4.1).

3.4. GGA evolution dynamics

Feature selection is performed with the GGA. All possible 92 predicting variables are assigned to one of the different groups defined in the groups' part of the individual. Since each individual is divided into several groups, there are different approaches to calculate the fitness function (given by Eq. (4)), as all groups may be used in this calculation, just one, and so on. In this work we have analyzed two different evolution dynamics for the GGA:

1. Dynamics $\mathcal{D}_1$: the fitness function given by Eq. (4) is calculated for all groups in each individual, and the fitness value is assigned by choosing the minimum value for all the groups.
2. Dynamics $\mathcal{D}_2$: the fitness function is also given by Eq. (4), but in this case we choose to maximize the value of this equation for group G_1.

Note that $\mathcal{D}_1$ is the most intuitive way of the GGA evolution, where the individuals are selected according to the best fitness value obtained for one of their groups. On the other hand, $\mathcal{D}_2$ is completely different: in this case the idea is to concentrate those features that produce a poor performance of the regressor in a given group (group G_1 in this case). At the end of the evolution, the test value is obtained using the features that are not present in the first group of the best individual in the population. Note that $\mathcal{D}_2$ can be improved by eventually removing the worst features out of the total available ones after a number of generations (η). In order to do this, after η generations of the algorithm, we construct a ranking of those features that appear the most in group G_1 of all the individuals in the population. We then set a threshold (th) for the number of times a given feature appears in group G_1, and we remove those features that appear in the ranking over threshold th. The GGA is then re-initialized without considering those features that were removed in the previous step.

4. Experiments and results

This section refers the experiments run to assess the proposed algorithm, together with measures used to evaluate the accuracy of our approach, and to allow comparison with other global solar radiation prediction tools.

Note these experiments consider global solar radiation prediction only for daytime hours (average hourly data from 5 a.m. to 8 p.m.), as night hours present zero irradiance. In order to create training and test sets to evaluate the performance, each day has been divided into several blocks. For each block, a random hour is assigned to the test set, and the remaining ones, in increasing order, to the training set. Thus, guaranteeing that all blocks and all days are represented both in train and test. Finally, this procedure is carried out 10 times, and we then provide average values of error in test (10-fold cross validation).

4.1. Forecast accuracy measures

Following the guidelines given in Beyer et al. (2011), in order to assess forecast accuracy of global solar radiation, conventional metrics such as RMSE (Eq. (4)) and Pearson Correlation Coefficient, r^2, have been used in this paper.

Moreover, to facilitate comparisons between the forecasts developed in this work and other solar forecasts, another useful quality check is to analyze whether the proposed GGA–ELM model performs better or worse than a given reference model (Kleissl, 2013). Thus, forecast skill, s, defined in Eq. (5) has been also considered, where U stands for the uncertainty of solar availability, i.e. the forecasting error of the proposed GGA–ELM model, and V refers to the variability of solar irradiance. Taking into account that this variability can be attributed to cloud cover (mostly stochastic) and solar position (mostly deterministic), it can be referred as the standard deviation of the step-changes of the ratio of the measured solar irradiance to that of a clear-sky solar irradiance. An easier procedure to obtain the forecast skill, and yet a good estimate (Kleissl, 2013), is to consider that the ratio U/V can be approximated by the $RMSE_{prediction}/RMSE_{persistencemodel}$. Global solar radiation obtained with the persistence model at a point $\mathcal{P}$ and at a time t, will be referred as $\mathcal{I}_t^{Per}$.

$$s = 1 - \frac{U}{V} \tag{5}$$

4.2. SubProblem 1

SubProblem 1 evaluates what WRF model outputs (the predictive variables) are more useful in the global solar radiation prediction. To assess it, we have carried out several experiments to test the proposed hybrid GGA–ELM algorithm. First, the GGA will perform a feature selection out of the 92 possible atmospherical output variables considered. Then, the ELM will perform the global solar prediction using those features selected by the GGA in the test set.

The result for global solar radiation prediction with the ELM when no feature selection is performed (tested on all the available features) is $r^2 = 0.9283$ and $RMSE = 85.14\ \mathrm{W/m^2}$ (average of the 10-fold cross validation). Note that this is a baseline reference that should be outperformed by the proposed algorithm. Figs. 5 and 6 present the scatter plot and global solar prediction in time, where it can be seen that the prediction fits rather well to the field (measured) data. For a clearer representation, only the first 100 h of the test output are shown in all time graphs.

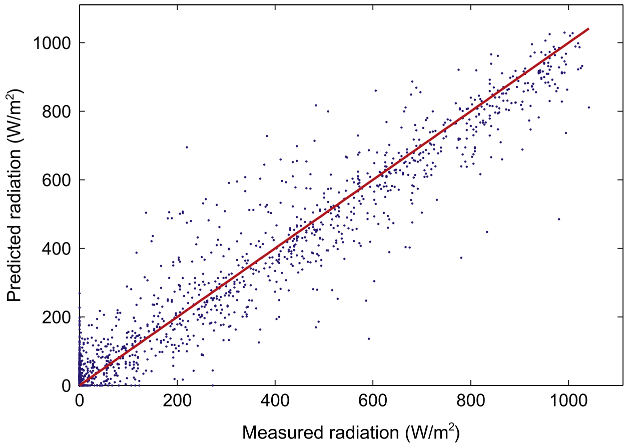


Fig. 5. Scatter plot of the global solar radiation prediction by the ELM without feature selection.

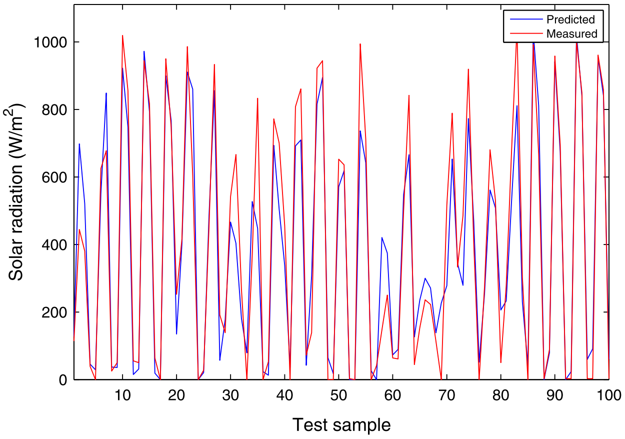


Fig. 6. Global solar radiation prediction in time by the ELM without feature selection.

Now, in order to test the hybrid GGA–ELM approach, several experiments were run showing the results presented in Table 2. Note that due to the ELM's output variability, when the hybrid GGA–ELM approach is tested, small increases in the RMSE may occur at certain generations. In order to reduce this variability, the ELM is run γ times ($\gamma = 10$) at each iteration (for both dynamics) and the average RMSE value is used as the individual's fitness value (the minimum RMSE of all possible groups in each individual for dynamics $\mathcal{D}_1$, or the maximum RMSE in the individual's group G_1 for dynamics $\mathcal{D}_2$). Therefore, every generation in $\mathcal{D}_1$, the algorithm runs γ ELMs for each group at each individual ($\gamma \cdot N_{Groups}$), while every generation in $\mathcal{D}_2$ only runs γ ELMs for group G_1 at each individual.

The first experiments compare the two crossover operators presented in Section 3.2: $\mathcal{C}_1$ (two-parents/one-child) and $\mathcal{C}_2$ (two-parents/two-children). It can be observed in Table 2 that $\mathcal{C}_1$ presents slightly worse predictions than $\mathcal{C}_2$. Observing the evolution along generations for the first crossover operator ($\mathcal{C}_1$), a continuous increase in the

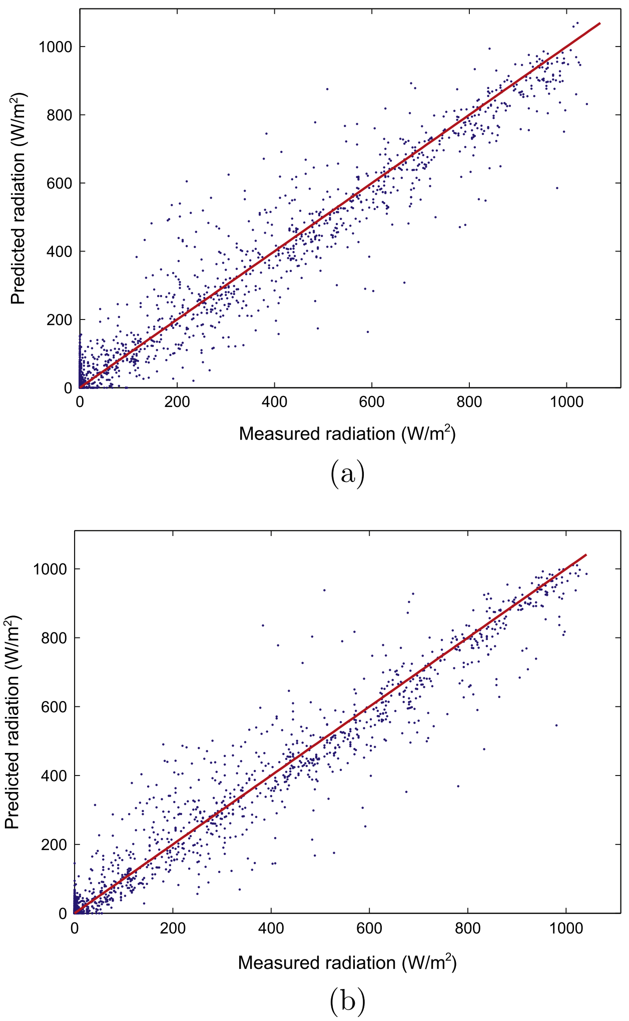
Table 2
Comparative results of the solar radiation prediction before and after feature selection with the GGA–ELM considering crossovers $\mathcal{C}_1$ and $\mathcal{C}_2$, and dynamics $\mathcal{D}_1$ and $\mathcal{D}_2$.

Experiments	$RMSE$ (W/m^2)	r^2
ELM (all features)	85.14	0.9283
GGA–ELM ($\mathcal{C}_1, \mathcal{D}_1$)	78.06	0.9382
GGA–ELM ($\mathcal{C}_1, \mathcal{D}_2$)	76.14	0.9406
GGA–ELM ($\mathcal{C}_2, \mathcal{D}_1$)	77.53	0.9401
GGA–ELM ($\mathcal{C}_2, \mathcal{D}_2$)	75.56	0.9415

number of groups was detected, and an upper bound in the number of groups to be created was imposed. Therefore, the number of groups was restricted to a maximum of 10, 15, 20 or 25, and any group created over this limit was destroyed and its items were randomly reallocated to existing groups. In spite of this consideration, $\mathcal{C}_2$ was found to outperform $\mathcal{C}_1$ in this problem.

Analyzing $\mathcal{C}_1$'s evolution along generations, the total number of groups tends to continuously increase, thus affecting improvement of the algorithm. Because of this, an upper bound in the maximum number of groups of 10, 15, 20 or 25 was set at each iteration, therefore destroying groups over this limit and randomly reallocating their items to existing groups. In spite of this consideration, $\mathcal{C}_2$ was found to outperform $\mathcal{C}_1$ in this problem.

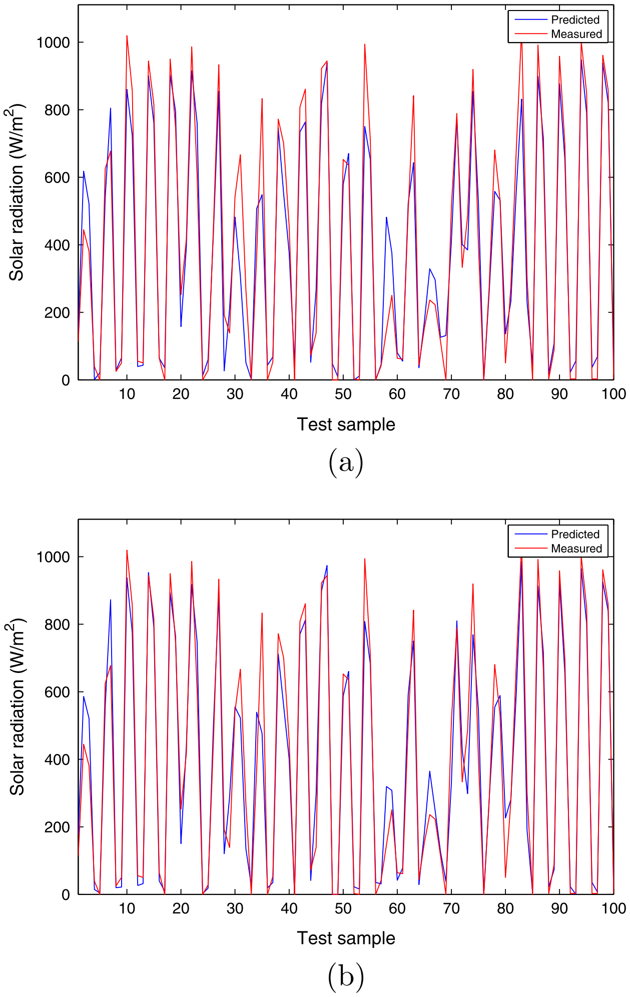
The last experiments compare the two different dynamics introduced in Section 3.4. The first one, $\mathcal{D}_1$, computes the fitness function for all groups in each individual, and the minimum value for all groups is set as the individual's fitness value. The second one, $\mathcal{D}_2$, maximizes each individual's first group fitness function and assigns it as the individual's fitness value. Therefore, $\mathcal{D}_1$ must run the ELMs on all groups in each individual, while $\mathcal{D}_2$ only



(a)

(b)

Fig. 7. Scatter plot of the global solar radiation prediction after feature selection with $\mathcal{C}_2$ crossover operator, and following dynamics: (a) $\mathcal{D}_1$; (b) $\mathcal{D}_2$.



(a)

(b)

Fig. 8. Global solar radiation prediction in time after feature selection with $\mathcal{C}_2$ crossover operator, and following dynamics: (a) $\mathcal{D}_1$; (b) $\mathcal{D}_2$.

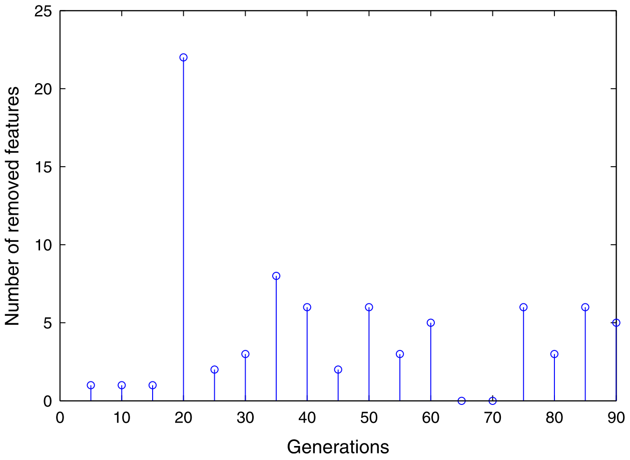


Fig. 9. Feature removal along generations with $\mathcal{C}_2$ crossover operator, and following dynamics $\mathcal{D}_2$.

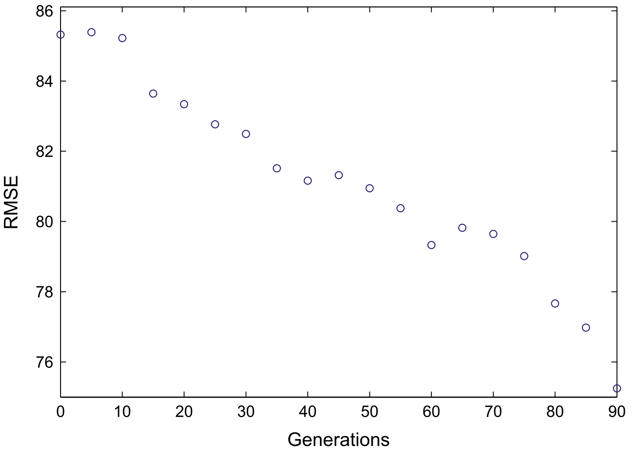


Fig. 10. GGA–ELM's performance along generations when $\eta = 5$ for $\mathcal{C}_2$ crossover operator, and following dynamics $\mathcal{D}_2$. Note that after η generations, several features are removed.

Table 3
Global solar radiation prediction for a 1, 2, and 3 h ahead forecasting, after performing a feature selection with the GGA–ELM (considering crossover $\mathcal{C}_2$ and dynamics $\mathcal{D}_2$).

GGA–ELM ($\mathcal{C}_2, \mathcal{D}_2$) Forecast horizon: $t + x$	*RMSE* (W/m^2)	r^2	s (ref: $\mathcal{I}_t^{Per}$)
$x = 1$ h	111.76	0.8693	13%
$x = 2$ h	165.86	0.7173	30%
$x = 3$ h	200.36	0.5900	40%

obtains G_1's fitness function. Moreover, $\mathcal{D}_1$ always performs the ELMs on the initial number of features, while $\mathcal{D}_2$ removes several features every $\eta = 5$ generations. This applied criterion requires that those features in G_1 that appear at least in 40% of the individuals are removed, and evolution continues with a smaller feature population. Let us focus on the second crossover operator (as it showed the best results), it can be seen that the RMSE decreases from 77.53 W/m^2, when applying $\mathcal{D}_1$, to 75.56 W/m^2 for $\mathcal{D}_2$, and r^2 increases from 0.9401 to 0.9415, respectively. Figs. 7 and 8 present, respectively, the scatter plot and the solar prediction in time for both dynamics. Once again, it can be seen that the prediction fits rather well to the measured data.

Analyzing the results of the 10-fold cross validation in the best experiment (crossover $\mathcal{C}_2$ and dynamics $\mathcal{D}_2$), it can be determined that the key predictive variables (features) are the OLR, CLDFRA (at a pressure level corresponding to 700 hPa), QVAPOR (at a pressure level corresponding to 700 hPa), and T (at a pressure level corresponding to 500 hPa), all for the WRF output point located at (39° 51′N, 4° 01′W). Other five to nine less important features complete the different GGA–ELM's solutions.

Finally, Fig. 9 shows the amount of features removed after each η generations, and it can be seen that sometimes no features are removed. Moreover, Fig. 10 presents the GGA–ELM's performance, where it can be observed that right after several characteristics are removed, the RMSE may increase, but on the long run, better results are obtained.

4.3. SubProblem 2

SubProblem 2 analyzes the prediction performance of the proposed GGA–ELM approach for a 1, 2, and 3 h ahead forecasting, considering only the best algorithms' configuration found in the previous subproblem (i.e. $\mathcal{C}_2$ crossover operator, and dynamics $\mathcal{D}_2$). Note that in this subproblem only the outputs of the WRF model are used as predictive variables. Table 3 presents the results obtained for this experiment, as well as the forecasting skill of the proposed GGA–ELM algorithm at point $\mathcal{P}$.

Table 4
Global solar radiation prediction for a 1 and 2 h ahead forecasting ($\hat{\mathcal{I}}_{t+x}$). Input variables include those features selected with the GGA–ELM approach at SubProblem 1 (crossover $\mathcal{C}_2$ and dynamics $\mathcal{D}_2$), together with past station's measurements ($\mathcal{I}_{t-z}$), ranging from $z = 0$ to $z = 5$.

	Predictive variables: ELM (all features) $+\mathcal{I}_{t-z}$			Predictive variables: GGA–ELM's best features $+\mathcal{I}_{t-z}$		
	RMSE (W/m^2)	r^2	s (ref: $\mathcal{I}_{t+x}^{Per}$)	*RMSE* (W/m^2)	r^2	s (ref: $\mathcal{I}_{t+x}^{Per}$)
Forecast horizon: x = 1						
$z = 0$	106.62	0.8838	17%	100.12	0.8970	22%
$z = 0, 1$	101.82	0.8935	21%	82.37	0.9299	36%
$z = 0$ to 2	95.43	0.9074	26%	78.67	0.9366	39%
$z = 0$ to 3	92.16	0.9141	28%	78.28	0.9377	39%
$z = 0$ to 4	90.37	0.9182	30%	76.86	0.9405	40%
$z = 0$ to 5	90.96	0.9166	29%	76.53	0.9407	41%
Forecast horizon: x = 2						
$z = 0$	171.59	0.6921	27%	154.81	0.7317	35%
$z = 0, 1$	149.90	0.7652	37%	126.89	0.8311	46%
$z = 0$ to 2	133.88	0.8147	43%	116.64	0.8585	51%
$z = 0$ to 3	130.17	0.8264	45%	113.69	0.8670	52%
$z = 0$ to 4	128.08	0.8321	46%	113.46	0.8678	52%
$z = 0$ to 5	126.70	0.8340	46%	112.07	0.6899	53%

It can be seen that the longer the forecast horizon is, the worse the proposed algorithm predicts the global radiation (i.e. r^2 falls from 0.8693 to 0.5900). On the other hand, the forecast skill shows that the GGA–ELM outperforms the persistence model as the forecast horizon increases (FS from 13% to 40%).

4.4. SubProblem 3

SubProblem 3 is also a forecasting experiment that analyzes prediction performance for a time horizon $t+x, x=0,\ldots,\mathcal{X}$. In this subproblem, objective global solar radiation data measured in Toledo's station for times $t-z, z=0,\ldots,\mathcal{Z}$, are included as predictive variables. Note that in this case, as in SubProblem 2, only the best algorithms' configuration (i.e. $\mathcal{C}_2$ crossover operator, and dynamics $\mathcal{D}_2$) and best features found in SubProblem 1 have been analyzed.

Table 4 presents the results obtained for a one and two hours ahead prediction, when past and known global solar radiation data are included. Experiments considering station measurements as input variables (ranging from one past value ($z=0$) to six past values ($z=0$ to 5)) are shown. For comparison purposes, forecast skill (s) is included.

It can be seen that, by performing a feature selection with the proposed approach, the prediction skill improves over the use of all 92 WRF predictive variables. It is important to highlight that when solar radiation from the previous hours are also considered as input, performance increases as well. Moreover, increase due to previous data seems to contribute more than any other inputs. In all cases the GGA–ELM's skill is better than the persistence model.

5. Conclusions

In this paper we have presented a novel hybrid Grouping Genetic Algorithm–Extreme Learning Machine (GGA–ELM) approach for accurate global solar radiation prediction problems. The GGA is included to obtain a reduced number of features for the prediction, and the ELM is used as a fast predictor for the solar radiation. The outputs from a numerical weather model (WRF) are used as input features for the ELM, to be selected by the GGA. A real solar radiation prediction problem for Toledo's radiometric observatory (Spain) has been tackled to show the goodness of the proposed approach.

Three subproblems have been analyzed then: first, in SubProblem 1, the prediction system proposed only uses the output of the WRF as inputs, without any other additional information to do the prediction. This case consists of a downscaling of the global solar radiation prediction to a point of interest. In this first subproblem we have introduced and tested different refinements to the GGA–ELM to improve the feature selection and prediction capabilities of the system: (1) two different GGA crossover operators, and (2) two different dynamics for the algorithm, one implying the ELM's error minimization of any group of the GGA, and the second one implying the ELM's error maximization for a specific group of the GGA, followed by the removal and re-initialization of the algorithm afterwards. For this first subproblem, we have found out that the best algorithm's configuration consists of a two-parents/two-children crossover plus the maximization, removal and re-initialization dynamics, which obtains the best results in terms of different error measures. This algorithm's configuration leads to a best solution with only 9 predictive features out of the initial 92.

The second subproblem is a prediction problem, where we have tried to predict the solar radiation at the point of interest at different time tags $t+x$, but again using predictive variables from the WRF only. In this case, the longer the forecast horizon, the better the GGA–ELM's performance is, in terms of different error measures and forecast skill. Finally, we have tackled a complete prediction problem by including previous values of measured solar radiation (as features for the ELM) plus the predictive variables from the WRF. We have proven that the inclusion of these previous radiation measures significatively improves the forecast skill with respect to persistence model.

Acknowledgement

This work has been partially supported by the project TIN2014-54583-C2-2-R of the Spanish Ministerial Commission of Science and Technology (MICYT), and by the Comunidad Autónoma de Madrid, under project number S2013ICE-2933_02. We would like to thank as well AEMET for providing the observed atmospherical data used in this work.

References

Agustín-Blas, L.E., Salcedo-Sanz, S., Vidales, P., Urueta, G., Portilla-Figueras, J.A., 2011. Near optimal citywide WiFi network deployment using a hybrid grouping genetic algorithm. Expert Syst. Appl. 38 (8), 9543–9556.

Alharbi, M.A., 2013. Daily Global Solar Radiation Forecasting using ANN and Extreme Learning Machines: A Case Study in Saudi Arabia (Master of Applied Science Thesis). Dalhousie University, Halifax, Nova Scotia.

Behrang, M.A., Assareh, E., Ghanbarzadeh, A., Noghrehabadi, A.R., 2010. The potential of different artificial neural network (ANN) techniques in daily global solar radiation modeling based on meteorological data. Sol. Energy 84 (8), 1468–1480.

Benghanem, M., Mellit, A., 2010. Radial basis function network-based prediction of global solar radiation data: application for sizing of a stand-alone photovoltaic system at Al-Madinah, Saudi Arabia. Energy 35 (9), 3751–3762.

Beyer, H.G., Polo Martinez, J., Suri, M., et al., 2011. Deliverable 1.1.3. Report on Benchmarking of Radiation Products. Report under contract no. 038665 of MESoR. <http://www.mesor.net/deliverables.html>.

Bhardwaj, S., Sharma, V., Srivastava, S., Sastry, O.S., Bandyopadhyay, B., Chandel, S.S., Gupta, J.R., 2013. Estimation of solar radiation using a combination of Hidden Markov Model and generalized Fuzzy model. Sol. Energy 93, 43–54.

Bilgili, M., Ozoren, M., 2011. Daily total global solar radiation modeling from several meteorological data. Meteorol. Atmos. Phys. 112 (3–4), 125–138.

Blum, A., Langley, P., 1997. Selection of relevant features and examples in machine learning. Artif. Intell. 97, 245–271.

Brown, E.C., Sumichrast, R.T., 2005. Evaluating performance advantages of grouping genetic algorithms. Eng. Appl. Artif. Intell. 18 (1), 1–12.

Carvalho, D., Rocha, A., Gómez-Gesteira, M., Silva Santos, C., 2014. Sensitivity of the WRF model wind simulation and wind energy production estimates to planetary boundary layer parameterizations for onshore and offshore areas in the Iberian Peninsula. Appl. Energy 135, 234–246.

Chen, J.L., Liu, H.B., Wu, W., Xie, D.T., 2011. Estimation of monthly solar radiation from measured temperatures using support vector machines – a case study. Renew. Energy 36, 413–420.

Coimbra, C., Kleissl, J., Marquez, R., 2013. Overview of solar forecasting methods and a metric for accuracy evaluation. In: Kleissl, J. (Ed.), Solar Resource Assessment and Forecasting. Elsevier, Waltham, Massachusetts.

Diagne, M., David, M., Boland, J., 2014. Post-processing of solar irradiance forecasts from WRF model at Reunion Island. Sol. Energy 105, 99–108.

Dong, H., Yang, L., Zhang, S., Li, Y., 2014. Improved prediction approach on solar irradiance of photovoltaic Power Station. TELKOMNIKA Indones. J. Electr. Eng. 12 (3), 1720–1726.

Dorvlo, A.S., Jervase, J.A., Al-Lawati, A., 2002. Solar radiation estimation using artificial neural networks. Appl. Energy 71 (4), 307–319.

Falkenauer, E., 1992. The grouping genetic algorithm–widening the scope of the GAs. Belg. J. Oper. Res. Stat. Comput. Sci. 33, 79–102.

Falkenauer, E., 1998. Genetic Algorithms and Grouping Problems. Wiley, New York.

Fu, C.-L., Cheng, H.-Y., 2013. Predicting solar irradiance with all-sky image features via regression. Sol. Energy 97, 537–550.

Giannaros, T.M., Kotroni, V., Lagouvardos, K., 2015. Predicting lightning activity in Greece with the weather research and forecasting (WRF) model. Atmos. Res. 156, 1–13.

Hocaoglu, F.O., Gerek, O.N., Kurban, M., 2008. Hourly solar radiation forecasting using optimal coefficient 2-D linear filters and feed-forward neural networks. Sol. Energy 82, 714–726.

Huang, G.B., Zhu, Q.Y., Siew, C.K., 2006. Extreme learning machine: theory and applications. Neurocomputing 70 (1), 489–501.

Huang, G.B., Chen, L., Siew, C.K., 2006. Universal approximation using incremental constructive feedforward networks with random hidden nodes. IEEE Trans. Neural Networks 17 (4), 879–892.

Huang, G.B., Wang, D.H., Lan, Y., 2011. Extreme learning machines: a survey. Int. J. Mach. Learn. Cybern. 2 (2), 107–122.

Huang, G.B., Zhou, H., Ding, X., Zhang, R., 2012. Extreme learning machine for regression and multi-class classification. IEEE Trans. Syst. Man Cybern. Part B 42 (2), 513–529.

Inman, R.H., Pedro, H.T., Coimbra, C.F., 2013. Solar forecasting methods for renewable energy integration. Prog. Energy Combust. Sci. 39 (6), 535–576, December 2013.

James, T., Brown, E.C., Keeling, K.B., 2007a. A hybrid grouping genetic algorithm for the cell formation problem. Comput. Oper. Res. 34, 2059–2079.

James, T., Vroblefski, M., Nottingham, Q., 2007b. A hybrid grouping genetic algorithm for the registration area planning problem. Comput. Commun. 30 (10), 2180–2190.

Ji, W., Chee, K.C., 2011. Prediction of hourly solar radiation using a novel hybrid model of ARMA and TDNN. Sol. Energy 85 (5), 808–817.

Kalogirou, S.A., 2014. Designing and modeling solar energy systems. Solar Energy Engineering. second ed., pp. 583–699, chapter 11.

Khatib, T., Mohamed, A., Sopian, K., 2012. A review of solar energy modeling techniques. Renew. Sustain. Energy Rev. 16, 2864–2869.

Kleissl, J. (Ed.), 2013. Solar Energy Forecasting and Resource Assessment. Academic Press.

Kohavi, R., John, G.H., 1997. Wrappers for features subset selection. Int. J. Digit. Libr. 1 (108–121).

López, G., Batlles, B., Tovar-Pescador, J., 2005. Selection of input parameters to model direct solar irradiance by using artificial neural networks. Energy 30, 1675–1684.

Mellit, A., Kalogirou, S.A., 2008. Artificial intelligence techniques for photovoltaic applications: a review. Prog. Energy Combust. Sci. 34 (5), 574–632.

Mohammadi, K., Shamshirband, S., Tong, C.W., Arif, M., Petkovic, D., Che, S., 2015. A new hybrid support vector machine-wavelet transform approach for estimation of horizontal global solar radiation. Energy Convers. Manage. 92, 162–171.

Mubiru, J., 2008. Predicting total solar irradiation values using artificial neural networks. Renew. Energy 33, 2329–2332.

Olatomiwa, L., Mekhilef, S., Shamshirband, S., Mohammadi, K., Petkovic, D., Ch, S., 2015. A support vector machine-firefly algorithm-based model for global solar radiation prediction. Sol. Energy 115, 632–644.

Paoli, C., Voyant, C., Muselli, M., Nivet, M.L., 2010. Forecasting of preprocessed daily solar radiation time series using neural networks. Sol. Energy 84, 2146–2160.

Rahimikoob, A., 2010. Estimating global solar radiation using artificial neural network and air temperature data in a semi-arid environment. Renew. Energy 35, 2131–2135.

Rehman, S., Mohandes, M., 2008. Artificial neural network estimation of global solar radiation using air temperature and relative humidity. Energy Policy 36 (2), 571–576.

Sahin, M., Kaya, Y., Uyar, M., Yidirim, S., 2014. Application of extreme learning machine for estimating solar radiation from satellite data. Int. J. Energy Res. 38 (2), 205–212.

Salcedo-Sanz, S., Prado-Cumplido, M., Pérez-Cruz, F., Bousoño-Calzón, C., 2002. Feature selection via genetic optimization. In: Proc. of the Artificial Neural Networks Conference (ICANN), pp. 547–552.

Salcedo-Sanz, S., M Pérez-Bellido, A., Ortiz-García, E.G., Portilla-Figueras, A., Prieto, L., 2009. Hybridizing the fifth generation mesoscale model with artificial neural networks for short-term wind speed prediction. Renew. Energy 34, 1451–1457.

Salcedo-Sanz, S., M Pérez-Bellido, A., Ortiz-García, E.G., Portilla-Figueras, A., Prieto, L., Correoso, F., 2009. Accurate short-term wind speed prediction by exploiting diversity in input data using banks of artificial neural networks. Neurocomputing 72 (4), 1336–1341.

Salcedo-Sanz, S., Casanova-Mateo, C., Pastor-Sánchez, A., Gallo-Marazuela, D., Labajo-Salazar, A., Portilla-Figueras, A., 2013. Direct solar radiation prediction based on soft-computing algorithms including novel predictive atmospheric variables. In: Intelligent Data Engineering and Automated Learning – IDEAL 2013. In: Lecture Notes in Computer Science, vol. 8206, pp. 318–325.

Salcedo-Sanz, S., Casanova-Mateo, C., Pastor-Sánchez, A., Sánchez-Girón, M., 2014. Daily global solar radiation prediction based on a hybrid Coral Reefs Optimization – Extreme Learning Machine approach. Sol. Energy 105, 91–98.

Salcedo-Sanz, S., Pastor-Sánchez, A., Prieto, L., Blanco-Aguilera, A., García-Herrera, R., 2014. Feature selection in wind speed prediction systems based on a hybrid coral reefs optimization – extreme learning machine approach. Energy Convers. Manage. 87, 10–18.

Senkal, O., Kuleli, T., 2009. Estimation of solar radiation over Turkey using artificial neural network and satellite data. Appl. Energy 86 (7–8), 1222–1228.

Skamarock, W.C., Klemp, J.B., Dudhia, J., Gill, D.O., Barker, D.M., Wang, W., Powers, J.G., 2005. A Description of the Advanced Research WRF Version 2. National Center for Atmospheric Reserach, Mesoscale and Microscale Meteorology Division, Technical Note.

Sozen, A., Arcakliogblu, E., Ozalp, M., 2004. Estimation of solar potential in Turkey by artificial neural networks using meteorological and geographical data. Energy Convers. Manage. 45, 3033–3052.

Voyant, C., Muselli, M., Paoli, C., Nivet, M.L., 2011. Optimization of an artificial neural network dedicated to the multivariate forecasting of daily global radiation. Energy 36 (1), 348–359.

Voyant, C., Muselli, M., Paoli, C., Nivet, M.L., 2013. Hybrid methodology for hourly global radiation forecasting in Mediterranean area. Renew. Energy 53, 1–11.

Wu, J., Chan, C.K., 2011. Prediction of hourly solar radiation using a novel hybrid model of ARMA and TDNN. Sol. Energy 85, 808–817.

Yacef, R., Benghanem, M., Mellit, A., 2012. Prediction of daily global solar irradiation data using Bayesian neural network: a comparative study. Renew. Energy 48, 146–154.

Yao, X., Liu, Y., Lin, G., 1999. Evolutionary programming made faster. IEEE Trans. Evol. Comput. 3 (2), 82–102.

Zeng, J., Qiao, W., 2013. Short-term solar power prediction using a support vector machine. Renew. Energy 52, 118–127.

Renewable Energy 111 (2017) 63–76

Contents lists available at ScienceDirect

Renewable Energy

journal homepage: www.elsevier.com/locate/renene

A CRO-species optimization scheme for robust global solar radiation statistical downscaling

S. Salcedo-Sanz [a], S. Jiménez-Fernández [a, *], A. Aybar-Ruiz [a], C. Casanova-Mateo [b, c], J. Sanz-Justo [b], R. García-Herrera [d, e]

[a] *Department of Signal Processing and Communications, Universidad de Alcalá, Alcalá de Henares, Spain*
[b] *LATUV: Remote Sensing Laboratory, Universidad de Valladolid, Valladolid, Spain*
[c] *Department of Civil Engineering: Construction, Infrastructure and Transport, Universidad Politécnica de Madrid, Madrid, Spain*
[d] *Department of Astrophysics and Atmospheric Sciences, Universidad Complutense de Madrid, Spain*
[e] *Department of Sedimentary Geology and Environmental Change, Instituto de Geociencias IGEO, UCM-CSIC, Madrid, Spain*

ARTICLE INFO

Article history:
Received 22 December 2016
Received in revised form
21 February 2017
Accepted 26 March 2017
Available online 28 March 2017

Keywords:
Coral reefs optimization algorithm
CRO with species
Global solar radiation
Solar energy
Extreme learning machines

ABSTRACT

This paper tackles the prediction of the global solar radiation (GSR) at a given point, using as predictive variables the outputs of a numerical weather model (the WRF meso-scale model) obtained at a different grid points. Prediction is obtained in this work using a Multilayer Perceptron (MLP) trained with Extreme Learning Machines (ELMs). Provided that the number of WRF outputs is vast, we propose the use of a Coral Reefs Optimization algorithm with species (CRO-SP) to obtain a reduced number of significant predictive variables, therefore improving the global solar radiation prediction attained without feature selection. The proposed system has been tested on real data from a radiometric station located at Toledo (Spain) and average best results of RMSE of 69.19 W/m^2 have been achieved, resulting in a 21.62% improvement over the average prediction without considering the CRO-SP for the feature selection.

1. Introduction

Solar energy is a clean and sustainable renewable source, with a high potential for significant growth in future years. Solar energy development is specially important in the Middle-East region, southern Europe and in the USA, places where the solar resource can be exploited all year around [1]. An important problem faced by this renewable resource is its integration in the grid system, because the energy produced by solar facilities is intrinsically stochastic due to the presence of clouds, atmospheric particles, dust, etc. In order to predict the solar production, an accurate global solar radiation (GSR) at the solar plant is needed, and this radiation depends completely on different atmospheric variables [2–4].

A vast set of different Machine Learning and Artificial Intelligence techniques has been applied for the prediction of global solar radiation [5]. Most of them use different machine learning techniques, with inputs based on meteorological and geographical parameters such as sunshine duration, air temperature, relative humidity, wind speed, wind direction, cloud cover, precipitation, etc [6,7]. According to [8] the design, control and operation of solar energy systems requires long-term series of meteorological data such as solar radiation, temperature, or wind data. Artificial Neural Networks (ANNs) are one of the most applied methods in solar radiation prediction problems. In Ref. [9] an exhaustive review on solar radiation prediction using ANNs is presented, describing forty two different researches, each one using as input variables several of the above-mentioned parameters to predict the solar radiation. In Ref. [10], the authors present a vast set of Artificial Intelligence techniques (ANN, Fuzzy logic, Genetic algorithms, Expert systems, etc.) applied to different photovoltaic applications: sizing of photovoltaic (PV) systems, modeling, simulation and control of PV systems or prediction of PV production using atmospheric or meteorological data. In some of the works cited, also meteorological and geographical parameters are used to increase the accuracy of the systems implemented, such as in Ref. [11], where ANNs are used with geographical parameters to estimate solar energy potential in Turkey. In Ref. [12], different combinations of input variables are considered and tested with Multi-Layer Perceptrons

* Corresponding author. Department of Signal Processing and Communications, Universidad de Alcalá, 28871, Alcalá de Henares, Madrid, Spain.
E-mail address: silvia.jimenez@uah.es (S. Jiménez-Fernández).

http://dx.doi.org/10.1016/j.renene.2017.03.079
0960-1481/© 2017 Elsevier Ltd. All rights reserved.

(MLP) and Radial Basis Function (RBF) neural networks. Moreover, results are then compared to conventional GSR prediction models, concluding that the ANN approaches perform better. In a similar research approach [13], carries out a comparison between MLP and RBF neural networks in a problem of solar radiation estimation. Experiments in eight stations in Oman show the good results obtained with the neural algorithms. The work in Ref. [14] presents the performance of ANN in solar energy prediction in of Kuwait. Two different training approaches, gradient descent and Levenberg-Marquart algorithm are tested in five different Kuwaiti locations. Recently, the use of Extreme Learning Machines (ELM) as a fast training method for ANNs has been applied due to the good GSR prediction results obtained. Sahin et al. [15] apply ELMs using satellite measures, concluding that the ELM model performs better than ANN with back-propagation in terms of GSR estimation and computational time. The ELM's performance in alternative radiation prediction systems is also described in Refs. [16–18].

Alternative strong regression algorithms have been applied in solar radiation prediction problems. Support Vector Regression (SVR) is one of these approaches, which has been successfully exploited in solar radiation prediction. In Ref. [19] the SVR algorithm has been mixed with a wavelet transform, in order to improve the performance of the former. Results in an Iranian coastal city have shown the performance of this hybrid proposal. Also with a hybrid algorithm involving SVRs [20], presents a firefly meta-heuristic with a SVR for global solar radiation prediction in different locations of Nigeria. In Ref. [21] solar irradiation mapping is tackled with SVR with exogenous data. Variable selection using a genetic algorithm is also considered in order to improve the performance of the SVR approach. In Ref. [22] a comparison between the performance of SVR and ANNs is carried out, in a problem of photovoltaic power generation. The work in Ref. [26] tackles a problem of global solar energy prediction with SVRs considering different prediction time horizons. Finally, in Refs. [23–25], a least-square SVR has been implemented using meteorological and atmospheric data as predictive variables. Specifically, the results obtained with the least-square SVR are compared to that of an auto-regressive neural network and a RBF neural network.

Bayesian methodology has also been applied in solar energy prediction problems. In Ref. [27] the description of a Bayesian methodology to determine the most relevant meteorological input parameters to an Artificial Neural Network (ANN) is introduced, concluding that there are relevant input parameters (clearness index and relative air mass) to estimate the direct irradiance. [28] also discusses the performance of Bayesian networks in global solar radiation prediction. Specifically, a comparison of GSR prediction performance between Bayesian networks, multi-layer perceptrons and empirical models is carried out. In Ref. [29], several clusters are formed and a prediction model is trained for each cluster to represent a different pattern in the stochastic component of the solar radiation, obtaining better results than ARMA or Time Delay Neural Networks. In Ref. [30], an approach for daily global solar irradiation prediction based on temporal Gaussian processes is discussed, improving the performance of a number of alternative regressors such as neural networks, SVR or regression trees.

Other approaches use all-sky or satellite images in order to obtain solar radiation prediction. In Ref. [31], the prediction of the solar radiation is based on different features extracted from all-sky images, such as the number of cloud pixels, frame difference, gradient magnitude, intensity level, accumulated intensity along the vertical line of the sun or the number of corners in the image. On the other hand [32], and [33] tackle the problem of solar radiation prediction from satellite images, in different locations in Turkey and Australia, respectively.

Hybrid approaches, i.e. algorithms which mix some kind of regression techniques with predictors from different sources have been quite used in solar energy prediction problems. In Ref. [34] solar irradiation in India is analyzed using a hybrid approach that combines Hidden Markov Models and generalized fuzzy models. According to Diagne et al. [35], forecasting of global horizontal irradiance (GHI) can be categorized according to the input variables used, that also determine the forecast horizon where they perform best. For instance, for time horizons from 4 to 6 h, the use of numerical weather prediction (NWP) models typically outperforms satellite-based predictions. Moreover, the use of the Weather Research and Forecasting (WRF) meso-scale model (a regional NWP model) hybridized with a Kalman filter reduces de GHI hour-ahead forecast relative root mean square error (rRMSE) from 35.20% to 22.33% [36]. In Ref. [37] a hybrid approach formed by Time Delay Neural Networks and Auto-regressive Moving Average (ARMA) models is proposed for short-term (hourly) solar radiation prediction in Singapore. A similar approach which also mixes ANN and ARMA models is proposed in Ref. [38]. In this case, the performance of the methodology is evaluated in different location of the Southern French coast and Corsica. In Ref. [39] ANNs are mixed with 2-D linear filters to obtain a hybrid approach for hourly solar radiation prediction. In Ref. [40] ANNs are hybridized with numerical weather prediction models for solving a problem of surface solar irradiance prediction in Brazil. In Ref. [41] a mixed approach formed by an ANN with wavelet transform is proposed for solving a problem of solar irradiance forecasting in 25 different locations of Singapore.

Finally, in Ref. [42] a Grouping Genetic Algorithm (GGA) is mixed with an ELM in a problem of solar radiation prediction at different time horizons. In that work, the GGA determines which WRF output variables best refine the GSR forecast (performed using the above-mentioned ELM approach). In Ref. [43] a hybrid ELM–Coral Reefs Optimization is proposed for solar radiation prediction in Southern Spain, in which the CRO is used to modify the ELM weights in order to improve its performance.

In this paper we propose the use of a novel hybrid approach to optimally predict the GSR at a given location. For this purpose, we use a vast set of meteorological and atmospheric variables provided by a numeric meso-scale model, the WRF, at different points close to the target location under study. Then, a Coral Reefs Optimization with species (CRO-SP) algorithm is used to determine the best subset of WRF variables that lead to a best forecast. This problem is known in the literature as *statistically downscaling* the GSR prediction of a meso-scale model to a given point [44]. Therefore, the ultimate goal of this approach is to evaluate what features (predictive variables) from the numerical model are useful for this forecast. The proposed CRO-SP algorithm is a novel co-evolution algorithm recently described in Ref. [45], especially well-suited for optimization problems with variable-length encodings. In our work, each species in the CRO-SP algorithm represents the use of a different number of WRF variables. Furthermore, the proposed complete solar prediction system (CRO-SP + ELM) will be tested with real data from a radiometric station in Toledo, Spain.

The structure of the rest of the paper is the following: next section presents the problem formulation. Section 3 describes the objective GSR dataset used to train and test the forecast system, and the outputs of the WRF meso-scale model used as input (predictive) variables. Section 4 describes the CRO-SP algorithm that will perform the feature selection and the ELM method for solar radiation downscaling. Section 5 shows the main results obtained with the proposed hybrid approach in a real solar radiation prediction problem at Toledo, Spain. Finally, Section 6 closes the paper giving some final conclusions and remarks.

2. Problem formulation

Let $\mathcal{I}_t$ be the global solar radiation in point $\mathcal{P}$ (a given location of the Earth's surface) at time t and let $\widehat{\mathcal{I}}_t$ be the prediction of the global solar radiation in $\mathcal{P}$ at the same time instant t.

Let $\mathscr{M}$ be a numerical meso-scale model and let $\mathcal{V}$ be the output of the model, at a time t at m different points of a grid. $\mathcal{V}$ consists of the prediction at time t of n different atmospheric variables, φ_{mn} ($m \in \{1..M\}$ and $n \in \{1..N\}$). Note that some or all of these variables may be registered at ground level ($l = 0$) or at different pressure levels ($l \in \{0...\mathscr{L}\}$). The output of $\mathscr{M}$ can be expressed as $\mathcal{V} = (\varphi_{11}, ..., \varphi_{1N}, \varphi_{21}, ..., \varphi_{2N}, ..., \varphi_{M1}, ..., \varphi_{MN})$, as shown in Fig. 1.

3. Objective variable data and predictive variables considered

This research tackles the global solar radiation downscaling from the WRF model outputs to a given point. In this work, $\mathcal{P}$ pinpoints at the radiometric station of Toledo, Spain (39° 53′N, 4° 02′W). Fig. 2 (a) shows the measuring station's location within the Iberian Peninsula. The predictive variables considered are the outputs of the WRF model at the two grid points ($M = 2$) closest to the station (in terms of the minimum euclidean distance) and located at (39° 51′N, 4° 01′W) and (40° 02′N, 4° 01′W), respectively, see Fig. 2(b).

3.1. Objective variable data

The objective variable data to train and test the algorithms corresponds to one year (from May 1st, 2013 to April 30th, 2014) of hourly global solar radiation data collected at Toledo's measuring station. This station is located at 39° 53′N latitude, 4° 02′W longitude and 515 MASL. Two constraints have been considered: 1) night hours present zero irradiance, and 2) a unique set of hours of interest, regardless of the season, is needed. Therefore, hourly data from 5 a.m. to 8 p.m. throughout the year is used in this analysis.

3.2. Predictive variables provided by the Weather Research and Forecasting model (WRF)

In this work, the meso-scale model $\mathscr{M}$ considered is the Weather Research and Forecasting (WRF) model [46] developed by the National Center for Atmospheric Research (NCAR), the National Centers for Environmental Prediction (NCEP), the Forecast Systems Laboratory (FSL), the Air Force Weather Agency (AFWA), the Naval Research Laboratory, the University of Oklahoma, and the Federal Aviation Administration (FAA) of the USA. WRF is a meso-scale numerical weather prediction system that has been used in a wide range of meteorological [47] and renewable energy applications [48]. Specifically, WRF model version 3.6 has been used in this research, and meteorological data have been calculated over a window ranging from 34° 34′N to 44° 28′N, and from 4° 25′W to 4° 23′E. In this window, the grid has 99 × 59, each grid element covers 15 × 30 km^2. Some of the atmospheric variables (outputs of the WRF) are calculated at a pressure level corresponding to the ground level ($l = 0$), and some others at the ground level and at 36 pressure levels above the ground in the vertical direction ($l = 1...36$).

The WRF model outputs considered in the study are the following:

- OLR: Top of atmosphere outgoing long-wave radiation (W/m^2).
- GLW: Downward long-wave flux at ground level (W/m^2).
- SWDOWN: Downward short-wave flux at ground level (W/m^2).
- u: Zonal wind component at different pressure levels (m/s).
- v: Meridional wind component at different pressure levels (m/s).
- w: Vertical wind component at different pressure levels (m/s).
- PSFC: Atmospheric pressure at ground level (hPa).

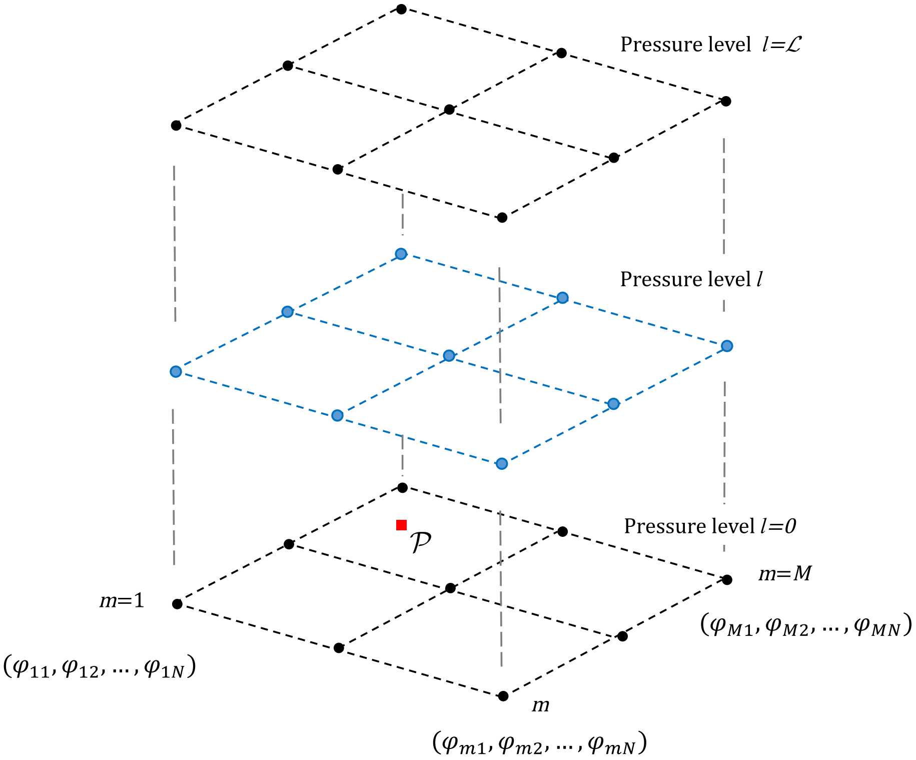
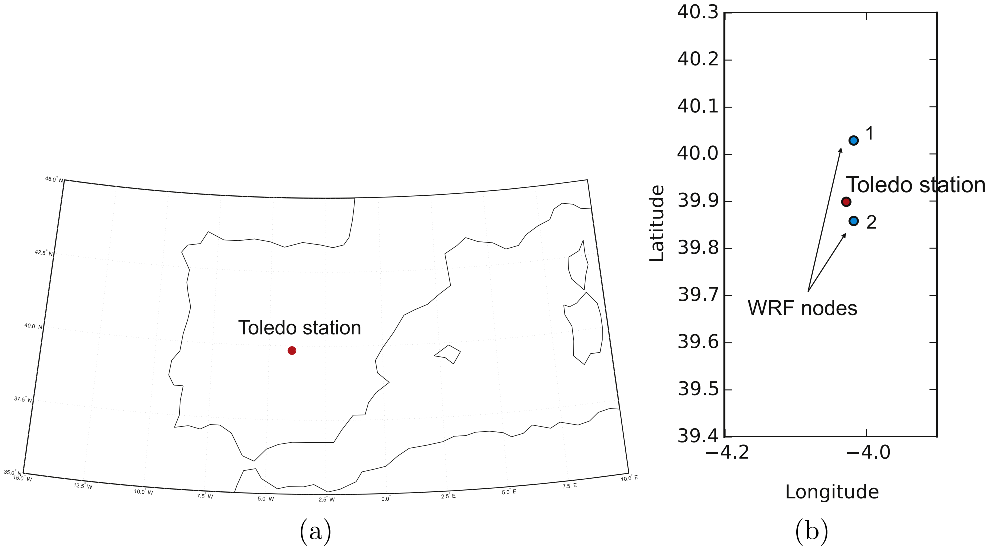

Pressure level $l=\mathcal{L}$
Pressure level l
Pressure level l=0
$\mathcal{P}$
m=1
$(\varphi_{11}, \varphi_{12}, ..., \varphi_{1N})$
m=M
$(\varphi_{M1}, \varphi_{M2}, ..., \varphi_{MN})$
m
$(\varphi_{m1}, \varphi_{m2}, ..., \varphi_{mN})$

Fig. 1. Global solar radiation prediction scheme used in this work.



(a) (b)

Fig. 2. Location of: (a) Toledo's measuring station in Spain and (b) the $M = 2$ WRF grid points considered for the downscaling.

- QVAPOR: Water vapor mixing ratio (in kg/kg). This variable is defined as the ratio of the mass of water vapor to the mass of dry air.
- TSK: Surface skin temperature (K).
- TH2: Potential temperature at 2 m above the ground (K).
- T′: Perturbation potential temperature (in K) at different pressure levels. The relationship between the perturbation potential temperature, T′, and the potential temperature, θ, is $\theta = T' + 300$.
- CLDFRA: Total cloudiness (fraction of clouds in each cell) at different pressure levels. Cloud fraction ranges from 0 (no clouds) to 1 (clouds in a spatial grid cell).

Table 1 shows the 58 variables analyzed for each grid point considered, indicating (when needed) the different pressure levels where they were obtained. Therefore, as two grid points ($M = 2$) have been examined, a total of 116 variables have been used in this work.

4. Methodology

In this work we use the Coral Reefs Optimization algorithm with Species to determine which set of WRF outputs obtains the best global solar radiation prediction. In Subsection 4.1, the basic CRO algorithm is introduced and in Subsection 4.2 a modification of the CRO including species (with co-evolution of variable-length encoding) is described to optimally tackle the feature selection. Next, in Subsection 4.3, the fitness (or health) function to be minimized by the abovementioned algorithm is presented, namely, the Extreme Learning Machine training method.

Table 1
Outputs of the WRF model used in the experiments as predictive variables (78 variables per point of the WRF model).

variable	pressure levels (hPa)
OLR	–
GLW	ground
SWDOWN	ground
u	ground, 850, 700, 500, 400, 300, 200, 100, 50
v	ground, 850, 700, 500, 400, 300, 200, 100, 50
w	ground, 850, 700, 500, 400, 300, 200, 100, 50
PSFC	ground
QVAPOR	ground, 850, 700, 500, 400, 300, 200, 100, 50
TSK	ground
TH2	ground
T′	ground, 850, 700, 500, 400, 300, 200, 100, 50
CLDFRA	ground, 850, 700, 500, 400, 300, 200

4.1. The basic Coral Reefs Optimization algorithm

The Coral Reefs Optimization (CRO) Algorithm [49] is an evolutionary algorithm based on the behavior of a coral reef that has been applied to a number of energy applications [50,51]. Let $\mathcal{R}$ be the reef represented by an $R_1 \times R_2$ grid, where each position (i,j) of $\mathcal{R}$ is able to allocate a coral or a colony of corals, $\mathcal{C}_{i,j}$. Therefore, a coral encodes a solution to the optimization problem under study. Namely, a subset of the WRF atmospheric variables that will be used to determine the GSR prediction: $\mathcal{C}_{i,j} = \{\phi_1, \phi_2, \ldots, \phi_K\}$, where each ϕ_k may be any of the atmospheric variables in $\mathcal{V}$.

The CRO algorithm first initializes some random positions of $\mathcal{R}$ with random corals and leaves some other positions empty. These holes in the reef are available to host new corals that will be able to freely settle and grow in later phases of the algorithm. The rate between free/occupied positions in $\mathcal{R}$ at the beginning of the algorithm is a parameter of the CRO algorithm denoted as ρ_0 $(0 < \rho_0 < 1)$.

The second phase simulates the processes of reproduction and reef formation. The different reproduction mechanisms available in nature are recreated by sequentially applying different operators. These behaviors are:

1. **External sexual reproduction or Broadcast Spawning.** Broadcast spawning consists of the following steps at each iteration k of the algorithm:
   1.a A random fraction of the existing corals is selected uniformly, turning these corals into broadcast spawners. The fraction of broadcast spawners with respect to the overall amount of existing corals in the reef will be denoted as F_b.
   1.b Several coral larvae are formed. To generate each new larva, two broadcast spawners are selected and a crossover operator or any other exploration strategy is applied. Note that once two corals have been selected to be the parents of a larva, they are not chosen anymore at iteration k for reproduction purposes. Corals' selection can be done randomly,

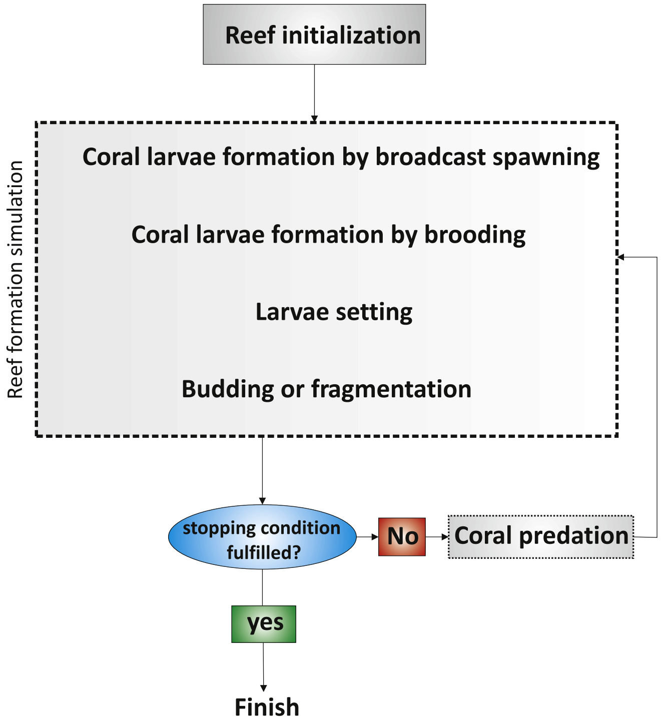

Reef initialization
Reef formation simulation
Coral larvae formation by broadcast spawning
Coral larvae formation by brooding
Larvae setting
Budding or fragmentation
stopping condition fulfilled?
No
Coral predation
yes
Finish

Fig. 3. Flowchart diagram of the original CRO algorithm.

uniformly, or using any fitness proportionate selection approach (e.g. roulette wheel).

2. **Internal sexual reproduction or Brooding.** Hermaphrodite corals reproduce by brooding. This reproduction is modeled by means of any kind of mutation mechanism and takes place on a fraction of corals of $1 - F_b$. A percentage P_i of the coral is mutated.
3. **Larvae setting.** Once the larvae are formed either through external or internal reproduction, they will try to set and grow in the reef. Each larva will randomly try to set in a position (i, j) of the reef and, if the location is free, it will set. If the location is already occupied, the new larva will set only if its health function (fitness) is better than that of the existing coral. Moreover, the CRO algorithm defines a parameter η that determines the maximum number of tries a larva can attempt at each iteration k.
4. **Asexual reproduction.** Corals reproduce asexually by budding or fragmentation. The CRO models this mechanism in the following way: the whole set of corals in the reef are sorted according to their level of health value (given by $f(\mathcal{C}_{i,j})$). Then, a small fraction (denoted as F_a) of the available corals are duplicated and mutated (with probability P_a) to provide variability, and try to settle in a different part of the reef as in Step 3.
5. **Depredation.** Corals may die during the reef's formation. Therefore, at the end of each reproduction iteration k, a small number of corals in the reef can be depredated, thus liberating space in the reef for next coral generation (iteration $k + 1$). The depredation operator is applied with a very small probability (P_d) to a fraction (F_d) of the corals in the reef with worse health.

Fig. 3 illustrates the flowchart diagram of the CRO algorithm, with the different CRO phases (reef initialization and reef formation), along with all the operators described above.

4.2. CRO algorithm with species: competitive co-evolution

The basic CRO can be improved to obtain stronger versions of the meta-heuristic, based on alternative processes that occur in coral reefs. We apply here the CRO-SP, a modification of the CRO that implements co-evolution to deal with optimization problems that present variable-length encodings. This advanced version of the algorithm was first described in Ref. [45] for a problem of optimal model selection.

Each *coral species* represents a different model (or its hyper-parameters) out of $\mathcal{N}$ possible models, and the concept of *model* is generic, so it can represent either a different encoding for the problem, a different way of obtain the health function, etc.

Bearing into mind that each coral individual is encoded as $\mathcal{C}_{i,j} = \{\phi_1, \phi_2, \ldots, \phi_K\}$, in this work, all corals belonging to a specific species are made up of the same number of WRF variables, K. For example, species $\mathcal{S}_1$ may represent solutions formed by $K = 10$ WRF variables, while species $\mathcal{S}_2$ may represent solutions formed by $K = 15$ predictive variables. Only corals of the same species can reproduce by external sexual reproduction described in Subsection 4.1. However, all species considered compete together in the larvae setting step, as only one health function is used for all species. Note that the competition among species will produce emerging behavior, so the best model (species) will eventually dominate and occupy the majority of spaces in the reef as the algorithm evolves. Algorithm 1 shows an outline of the CRO algorithm with species (CRO-SP).

Algorithm 1 Pseudo-code for CRO algorithm with species (CRO-SP)

```
Require: Valid values for CRO parameters.
Ensure: The best model out of $\mathcal{N}$ possible.
Algorithm initialization for $\mathcal{N}$ different species.
for each iteration $k$ of the CRO do
    Update values of influential variables: mortality probability and probability of asexual reproduction.
    Asexual reproduction (budding or fragmentation).
    External sexual reproduction (broadcast spawning, only same species can reproduce).
    Internal sexual reproduction (brooding).
    Settlement of new larvae (competition among species).
    Depredation.
    Evaluation of the new population in the reef (with the specific model given for each species).
end for
```

Table 2
CRO optimization parameters.

Phase	Parameter
Inicialization	Reef size $= 50 \times 40$ (2000 positions) $\mathcal{S}_i$, $i \in \{1..5\}$ (5 species) $\rho_0 = 0.75$ (1500 corals) $\rho_0^{\mathcal{S}_i} = 0.15$ (300 corals per species)
External sexual reproduction	$F_b = 0.70$ Random selection of broadcast spawners. Each possible coral must be broadcast spawner at least once per iteration k. New larva formation using 2-point crossover.
Internal sexual reproduction	$1 - Fb = 0.20$ $P_i = 0.30$
Larvae setting	$\eta = 3$ Identical corals are not allowed in the reef.
Asexual reproduction	$F_a = 0.05$ $P_a = 0.005$
Depredation	$F_d = 0.15$ $P_d = 0.25$ (it decreases with the number of iterations. At k_{max}, $P_d = 0$)
Stop criteria	$k_{max} = 300$ iterations.

It can be concluded that this behaviour increases the effectiveness of the algorithm, as different corals (individuals) belonging to different species compete against one another at each iteration of the algorithm. Traditional algorithms would evolve each experiment (species) separately and, at the end, compare them. Thus, resulting in longer processes. On the contrary, the main drawback of the proposed algorithm is its computational complexity, due to the competition at each iteration of all species present in the experiment.

4.3. *Extreme Learning Machines*

The above-mentioned CRO-SP is directly applicable to any problem of feature selection, since the number of features is usually a parameter to be set previously to the optimization process. In this problem of global solar radiation downscaling, feature selection is essential, as 116 predictive variables are considered. Thus, each CRO species is set to model a different number of features to be used in the optimization process, and the objective (health) function to be minimized is the downscaling error, given by a neural network trained with an Extreme Learning Machine algorithm.

The *Extreme Learning Machine* (ELM) is a training method for neural networks with the structure of multi-layer perceptrons. It obtains an extremely fast training by means of randomly setting the weights of the network's input layer, and then obtaining the weights of the output layer by using the inverse of the hidden-layer output matrix [52,53]. ELMs have shown excellent performance both in classification and regression applications [54]. For this reason, in this paper we have chosen the ELM as the regressor to determine the global solar radiation at $\mathcal{P}$.

In order to define the ELM process, consider a training set $\Gamma = \{(\mathbf{x}_i, \mathbf{y}_i) | \mathbf{x}_i \in \mathbb{R}^n, \mathbf{y}_i \in \mathbb{R}^m, i = 1, \cdots, N\}$, where $\mathbf{x}$ stands for the predictive variables, and $\mathbf{y}$ stands for the objective variable. It is also necessary to define an activation function for the neurons of the network, $a(x)$, and number of hidden nodes, γ. With these parameters, the ELM algorithms works as follows:

(1) Randomly assign the network's inputs weights $\mathbf{w}_i$ and bias b_i, $i = 1, \cdots, \gamma$.
(2) Calculate the hidden layer output matrix $\mathbf{H}$, as

$$\mathbf{H} = \begin{bmatrix} a(\mathbf{w}_1 \cdot \mathbf{x}_1 + b_1) & \cdots & a(\mathbf{w}_\gamma \cdot \mathbf{x}_1 + b_\gamma) \\ \vdots & \cdots & \vdots \\ a(\mathbf{w}_1 \cdot \mathbf{x}_N + b_1) & \cdots & a(\mathbf{w}_\gamma \cdot \mathbf{x}_N + b_\gamma) \end{bmatrix}_{N \times \gamma} \tag{1}$$

(3) Calculate the output weight vector β as

$$\beta = \mathbf{H}^\dagger \mathbf{Y}, \tag{2}$$

where $\mathbf{H}^\dagger$ stands for the Moore-Penrose inverse of matrix $\mathbf{H}$ [52], and $\mathbf{Y}$ is the training output vector, $\mathbf{Y} = [\mathbf{y}_1, \cdots, \mathbf{y}_N]^T$.

Finally, note that the number of hidden nodes (γ) is a free parameter of the ELM algorithm, so it must be estimated for a good performance of the algorithm. Usually, the solution for this problem is to scan a range of γ values and keeping the value which produces the best result.

4.4. *Fitness function*

The health or fitness function considered for each coral (individual) is obtained computing the Root Mean Square Error (RMSE) of the global solar radiation prediction, as shown in Eq. (3), where, again, $\mathcal{I}_t$ stands for the global solar radiation measured at a time t, $\widehat{\mathcal{I}_t}$ stands for the global solar radiation estimated by the ELM, and T stands for the number of samples in the test set.

$$RMSE = \sqrt{\frac{1}{T} \sum_{t=1}^{T} \left(\widehat{\mathcal{I}}_t - \mathcal{I}_t \right)^2} \tag{3}$$

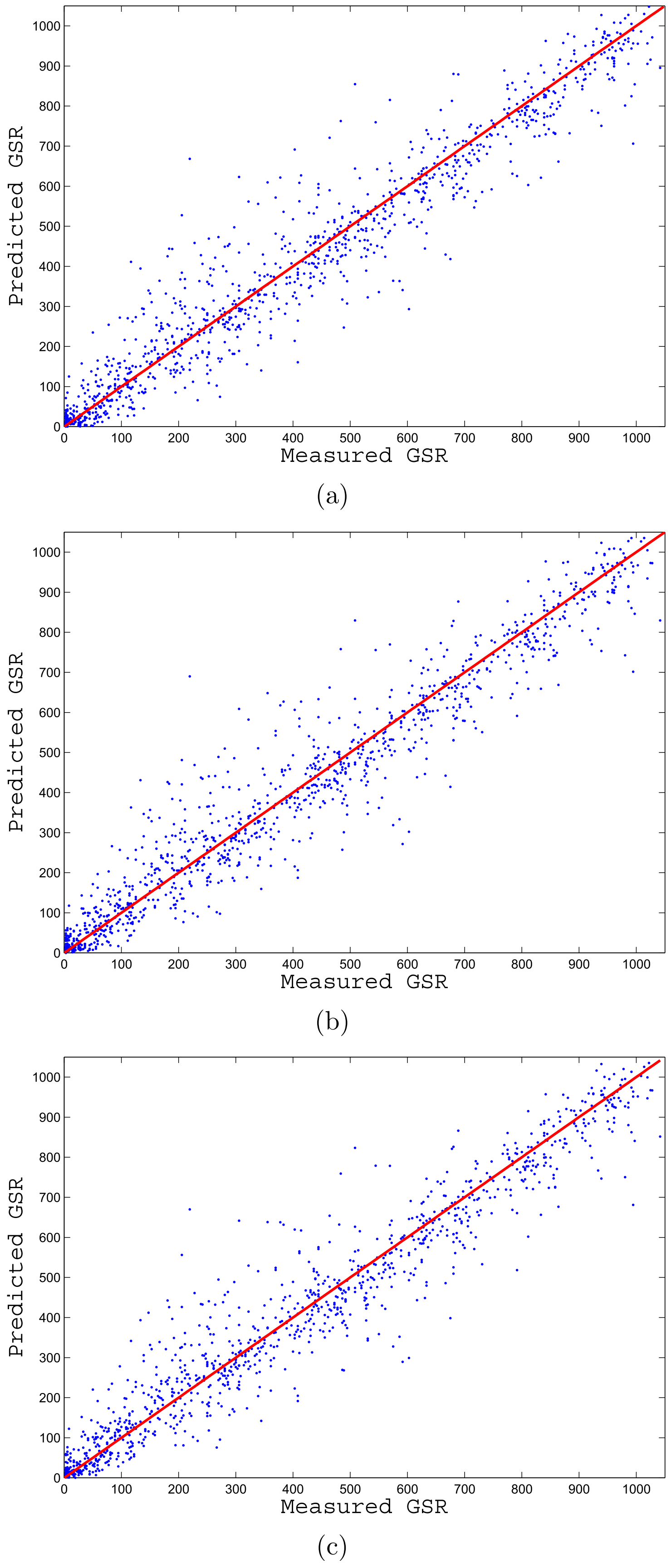
RMSE is used in this work instead of other validation metrics, because large forecast errors and outliers are weighted more strongly than smaller errors, as the latter are more tolerable in global solar radiation prediction [55,56].

5. Experiments and results

This section describes the experiments run in this work. Again, the aim is to predict the global solar radiation at a point P using as predictive variables the outputs of the WRF model obtained in two grid points close to P. The first step is to determine the best predictive variables (feature selection) and has been addressed

Table 3
Experiments run considering different species. Each *species* is represented by $\mathcal{S}_i$.

Experiment	Number of features per species					RMSE (W/m^2)		Best Species
	$\mathcal{S}_1$	$\mathcal{S}_2$	$\mathcal{S}_3$	$\mathcal{S}_4$	$\mathcal{S}_5$	Average	Best coral	
$\mathcal{E}_1$	10	20	30	40	50	69.50	68.21	10 features ($\mathcal{S}_1$)
$\mathcal{E}_2$	6	8	10	12	14	69.33	68.16	8 features ($\mathcal{S}_2$)
$\mathcal{E}_3$	7	8	9	10	11	69.16	68.03	8 features ($\mathcal{S}_2$)

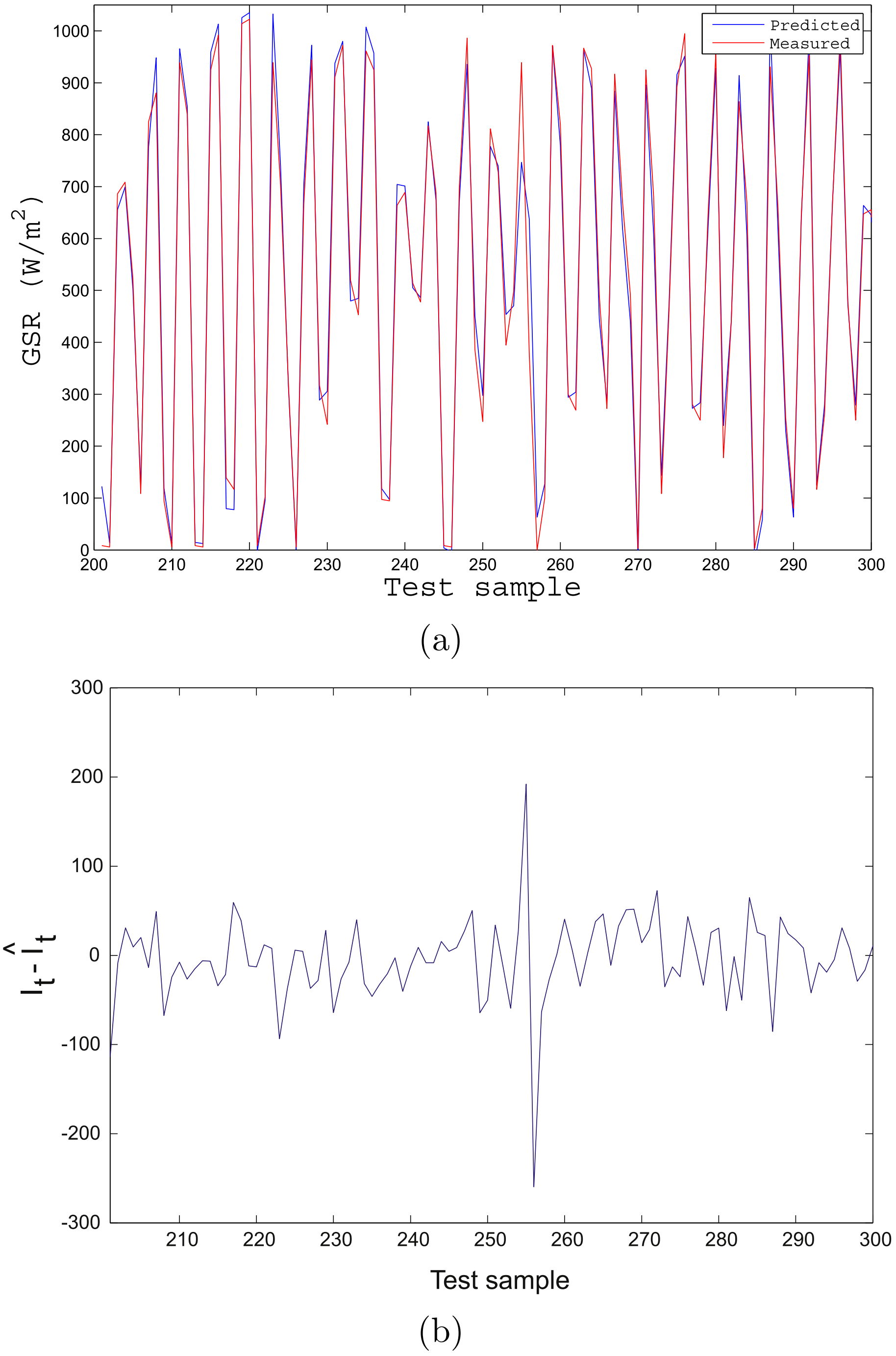


(a)

(b)

(c)

Fig. 4. Scatter plot of the global solar radiation: (a) Experiment $\mathcal{E}_1$, (b) Experiment $\mathcal{E}_2$ and (c) Experiment $\mathcal{E}_3$.



(a)

(b)

Fig. 5. Experiment $\mathcal{E}_3$. (a) Global solar radiation in time. (b) Deviation in time of the predicted GSR from the measured GSR. Note that only a random time frame of 100 samples is presented for clarity purposes.

implementing a CRO-SP algorithm with the parameters shown in Table 2.

To identify the best set and number of predictive variables, several experiments ($\mathcal{E}_i$, $i \in [1..3]$) have been run in a 10-fold cross validation scheme. Each CRO-SP experiment $\mathcal{E}_i$ consists of five subexperiments, each one of them analyzing a specific species $\mathcal{S}_i$ ($i \in [1..5]$), i.e., all corals belonging to one species have the same number of features. The co-evolution of these species leads quickly to a coral-reef colonized by the most suited corals. Once convergence is reached, the best coral in the reef belongs to a specific species and its health function stands for the RMSE value obtained in test. Note that to calculate the global solar radiation at each iteration, the ELM has been run 3 times and the health function value assigned to the coral is the average result obtained. Table 3 presents the results obtained for each experiment: the average value of error in test, the best coral's RMSE and its corresponding species (in terms of the number of predictive variables in that species).

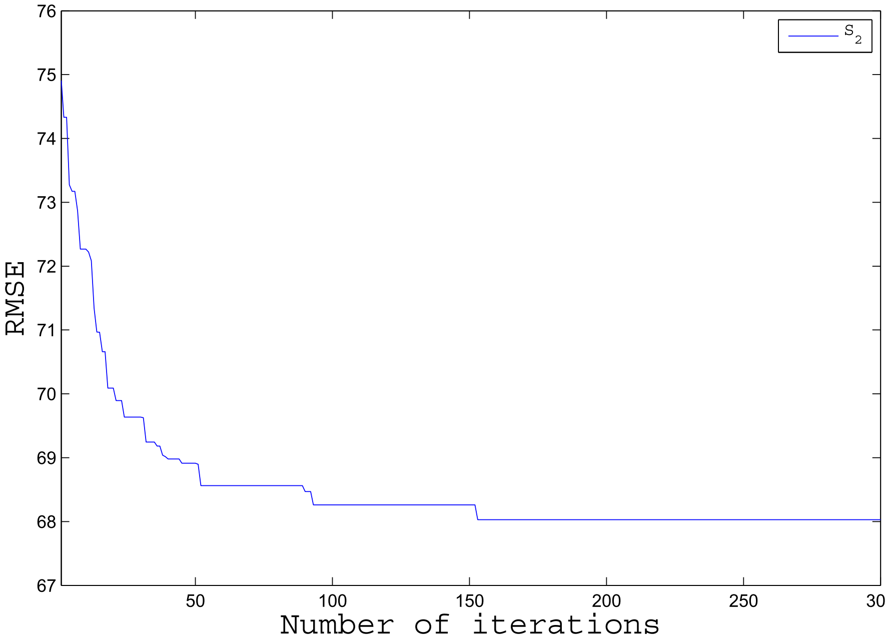


Fig. 6. Experiment $\mathcal{E}_3$. Evolution with the number of iterations of the best coral's RMSE. Note that the best coral belongs to species $\mathcal{S}_2$.

The first experiment run, $\mathcal{E}_1$, is meant to resolve the order of magnitude of the number of features to be considered (10, 20, 30, 40 or 50), and it can be observed that the best prediction is found using 10 variables (RMSE $= 68.21\ W/m^2$). Experiments $\mathcal{E}_2$ and $\mathcal{E}_3$ are used to refine the number of predictive variables to consider, both of them converging to best results when the species encode 8 variables. Fig. 4 presents the scatter plots for each experiments' best coral, showing the algorithm's good performance in all cases.

Fig. 5 presents the comparison in time between the measured and the predicted GSR for experiment $\mathcal{E}_3$, the best experiment, where it can be seen that the prediction follows rather well the field (target) data. Fig. 6 shows the evolution with the number of iterations of the best coral in this experiment. It corresponds to a coral encoding the use of 8 WRF variables and presents a final RMSE of $68.03\ W/m^2$ and a coefficient of determination $r^2 = 0.9531$.

It is interesting to analyze the behavior (evolution) of the

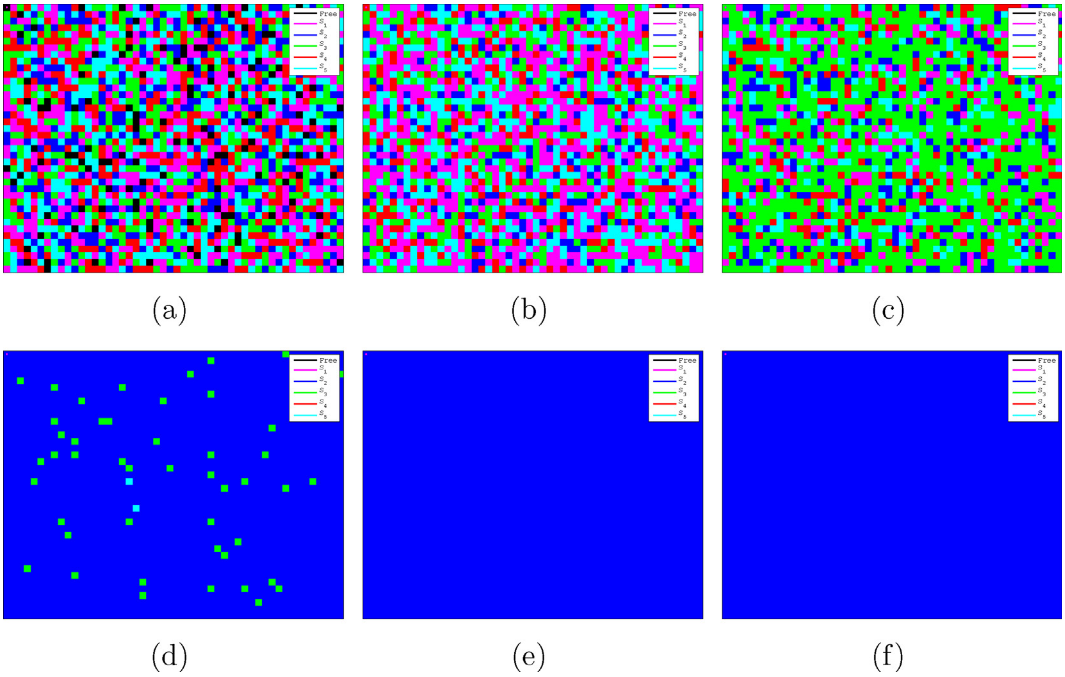


Fig. 7. Experiment $\mathcal{E}_3$. Evolution of the species present in the reef after a certain number of iterations (k): (a) $k = 1$, (b) $k = 10$, (c) $k = 25$, (d) $k = 50$, (e) $k = 150$, (f) $k = 300$.

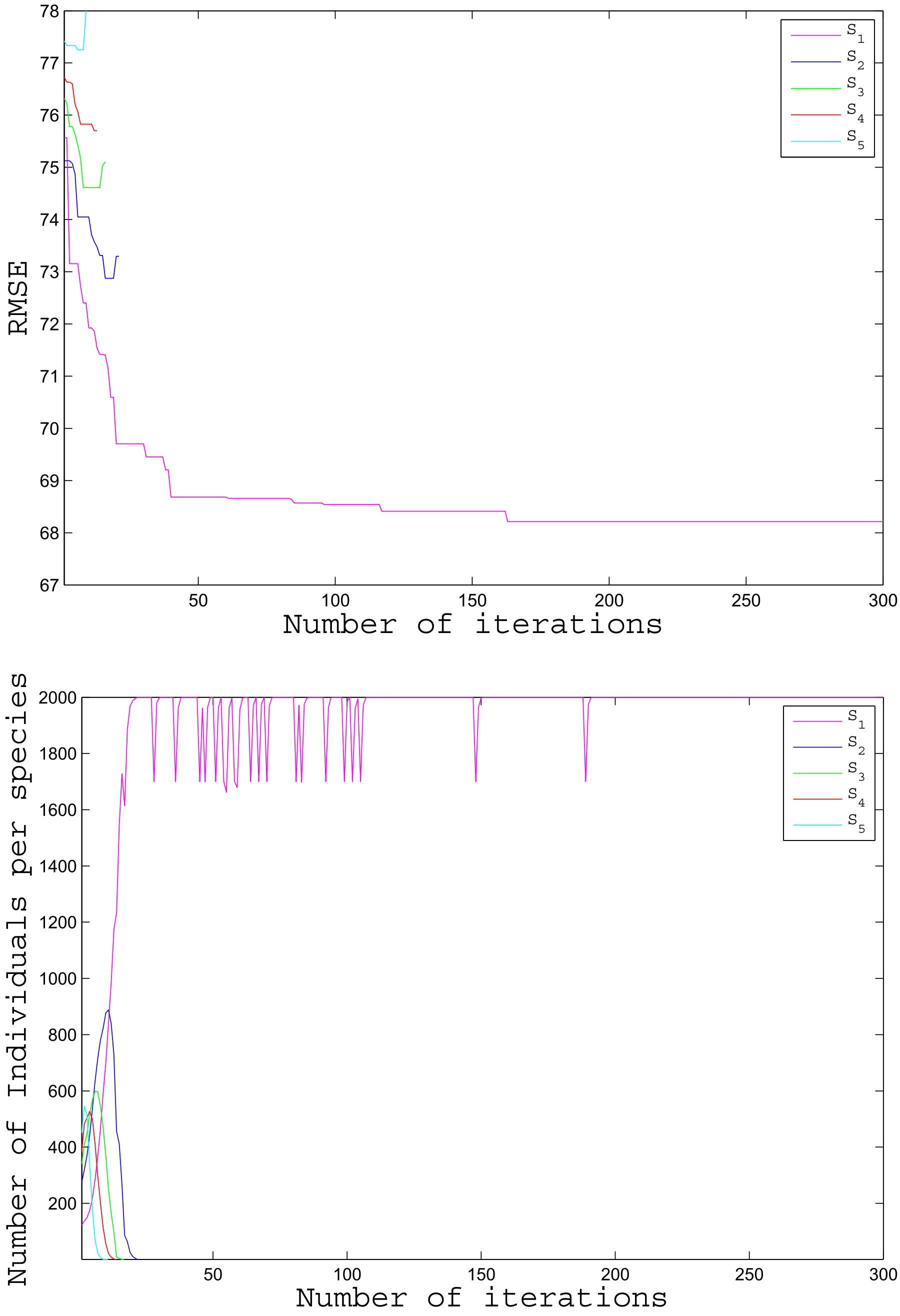


Fig. 8. Experiment $\mathcal{E}_1$. Evolution with the number of iterations of: (a) The RMSE of each species' best coral, and (b) The number of corals per species.

different species in the reef with the number of iterations. Fig. 7 shows this evolution for the best run of the best experiment, $\mathcal{E}_3$. In Fig. 7(a) the random initialization of the reef is presented, where the reader can see the positions occupied by each different species and the free positions available at the reef. As the number of iterations (k) increases (Fig. 7(b)-(f)), it can be observed that the worst-fitted species tend to die and are no longer present at the reef, as larvae from dominant species outperform them. Finally, when the stop criteria is reached, the reef is colonized by the best species which, in this particular experiment, is species $\mathcal{S}_2$ (corresponding to the use of 8 WRF variables for the prediction).

Figs. 8–10 show, for all experiments analyzed, the evolution with the number of iterations of two important characteristics. First, the root mean square error of each species' best coral, which is depicted in subfigures (a). It is clear that the RMSE decreases with the number of evolutions, but there is one exception: when a species is endangered (is being outperformed by the rest) its RMSE increases abruptly. Right after this occurs, the RMSE is interrupted, resulting in the disappearance of the worst-fitted species from the reef. Second, the number of corals present in each species is analyzed in Figs. 8(b), 9(b) and 10(b). It can be observed that, at some points, the number of corals in some species drops down. This is directly related to the occurrence of depredation phases. It is important to highlight that although in the depredation phase the species are decimated, the evolution

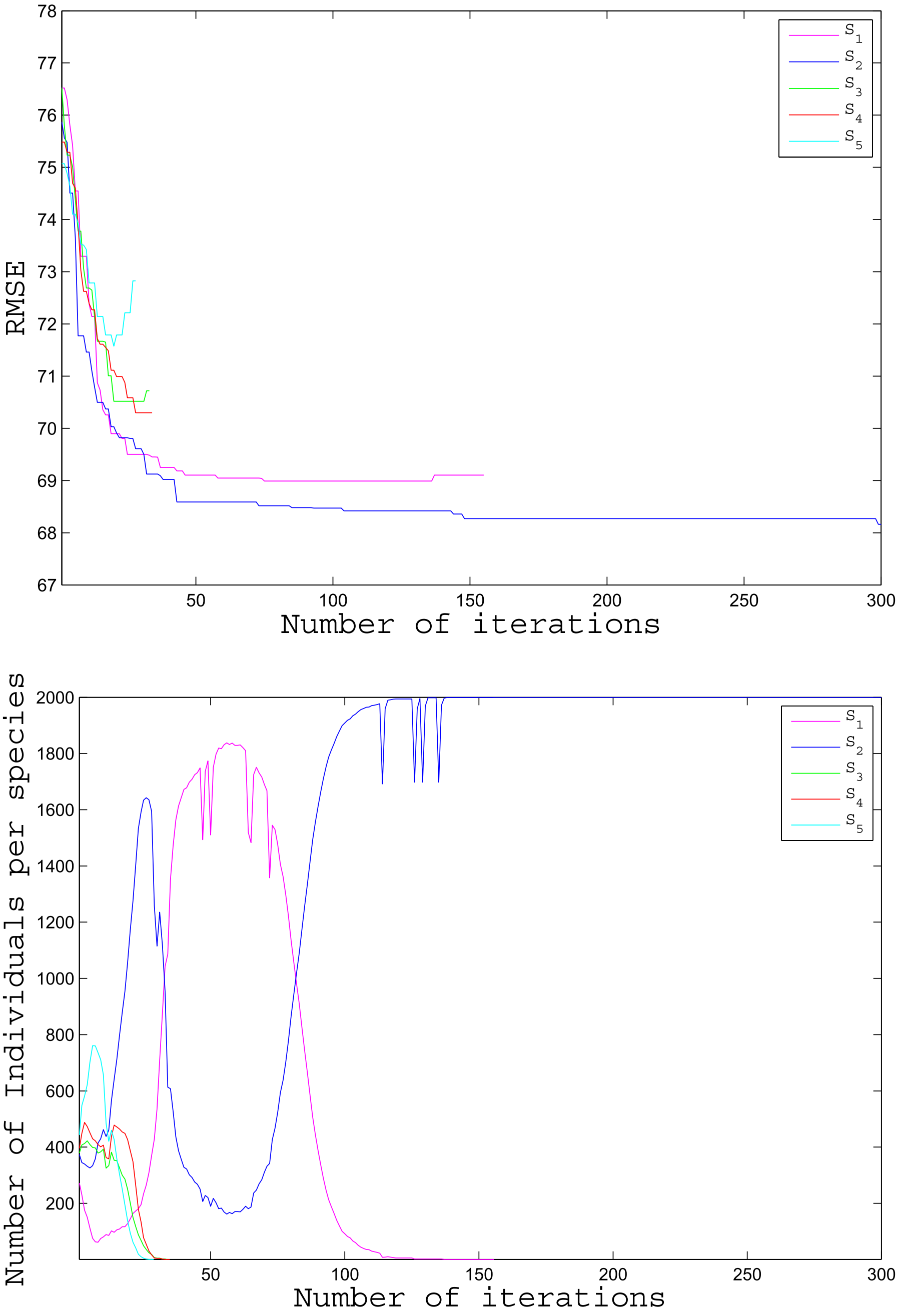


Fig. 9. Experiment $\mathcal{E}_2$. Evolution with the number of iterations of: (a) The RMSE of each species' best coral, and (b) The number of corals per species.

keeps recovering the best-fitted.

Next, Table 4 shows the name of the best coral's WRF outputs selected for each experiment. It can be seen that there are six variables: OLR, w at 400 hPa, CLDFRA at 200 hPa, T at 850 hPa and T at 400 hPa corresponding to the first grid point, and v at 500 hPa corresponding to the second grid point, present in all experiments' results. Therefore, we can conclude that these variables set the rough prediction while the other WRF outputs selected by the algorithm perform the refinement. Thus, for the third experiment, the RMSE using these 6 variables (over the same test sets) is 74.05 W/m^2 and 72.59 W/m^2, average and best values respectively. Once the refinement takes place, these RMSE values drop down to 69.16 W/m^2 and 68.03 W/m^2 respectively (as stated in Table 3).

Finally, in Table 5 the results are compared to those obtained with other techniques. First, the reader can see the GSR prediction using the 116 WRF variables (no feature selection) as inputs to the ELM. Then, feature selection is performed using three different techniques: a genetic algorithm, a grouping genetic algorithm (as described in Ref. [42]) and the proposed CRO-SP approach, and the variables chosen are used as the inputs to the ELM. It can be seen that the best results are obtained when the CRO with species is used.

6. Conclusions

This paper has tackled the global solar radiation prediction using

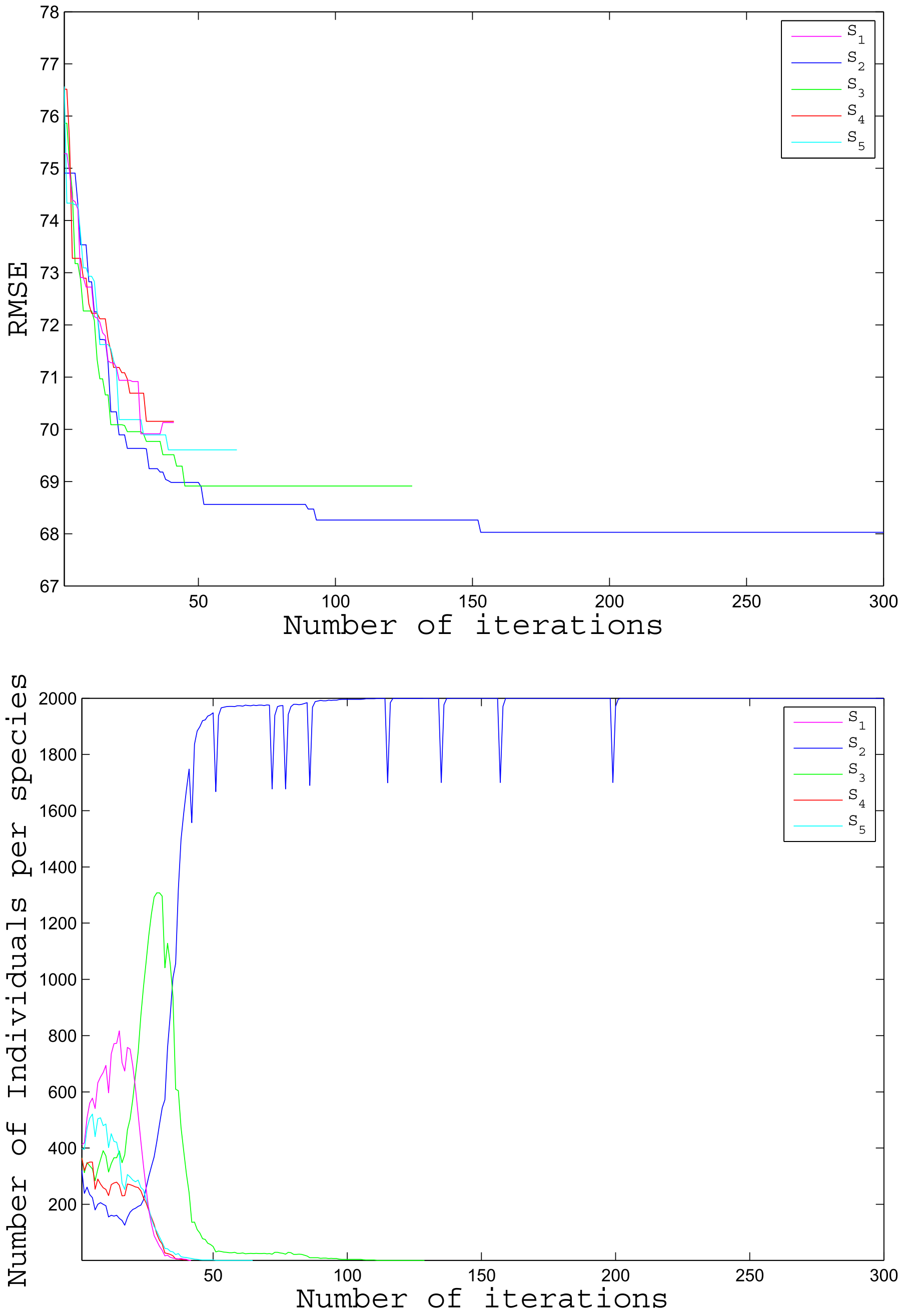


Fig. 10. Experiment $\mathcal{E}_3$. Evolution with the number of iterations of: (a) The RMSE of each species' best coral, and (b) The number of corals per species.

as predictive variables the outputs provided by a numerical weather model in a grid area over the target point (located at Toledo, Spain). The selection of the best predictive variables and the best grid points to be considered has been performed using the novel co-evolution algorithm *Coral Reefs Optimization algorithm with species* (CRO-SP), and the prediction has been obtained using a Multilayer Perceptron trained with Extreme Learning Machines (ELMs). The ultimate goal has been to evaluate what predictive variables from the numerical weather model (i.e. the WRF model) perform best. For this purpose, each species in the CRO-SP encodes a fixed and different number of variables to be analyzed and the best species comes out as a result of the co-evolution.

To determine the best set and number of predictive variables, three experiments have been run in a 10-fold cross validation scheme and the RMSE has been used as common measure. The experiment where 7, 8, 9, 10 and 11 variables are co-evolved (experiment $\mathcal{E}_3$), produces an average best result of 69.19 W/m^2 an a best result of 68.03 W/m^2, turning in a 21.62% and 22.03% improvement, respectively, over the average and best prediction without feature selection.

Table 4
Best predictive variables found for each experiment. Those variables present in all three experiments' results have been highlighted in bold face.

Experiment	Best WRF outputs selected	
	Grid point #1	Grid point #2
$\mathcal{E}_1$	**OLR**, ***w*(400 hPa)**, **CLDFRA (200 hPa)**, **T (850 hPa)**, **T (400 hPa)**, *u* (100 hPa)	***v*(500 hPa)**, PSFC, TH2, *u* (50 hPa)
$\mathcal{E}_2$	**OLR**, ***w*(400 hPa)**, **CLDFRA (200 hPa)**, **T (850 hPa)**, **T (400 hPa)**,	***v*(500 hPa)**, TH2 *w* (300 hPa)
$\mathcal{E}_3$	**OLR**, ***w*(400 hPa)**, **CLDFRA (200 hPa)**, **T (850 hPa)**, **T (400 hPa)**,	***v*(500 hPa)**, *u* (850 hPa) *u* (50 hPa)

Table 5
Comparison of the results obtained with other metaheuristic techniques.

Metaheuristic technique	RMSE (W/m^2)	
	Average	Best individual
No feature selection + ELM	88.24	87.25
Genetic algorithm + ELM	73.98	72.20
Grouping Genetic Algorithm + ELM [42]	74.73	73.66
CRO-SP + ELM	69.16	68.21

Acknowledgement

This work has been partially supported by the project TIN2014-54583-C2-2-R of the Spanish Ministerial Commission of Science and Technology (MICYT), and by the Comunidad Autónoma de Madrid, under project number S2013ICE-2933_02. We would like to thank as well AEMET for providing the observed atmospherical data used in this work.

References

[1] S.A. Kalogirou, Designing and Modeling Solar Energy Systems, Solar Energy Engineering (Second Edition), (Chapter 11), 2014, pp. 583–699.

[2] R.H. Inman, H.T. Pedro, C.F. Coimbra, Solar forecasting methods for renewable energy integration, Prog. Energy Combust. Sci. 39 (no. 6) (December 2013) 535–576.

[3] T. Khatib, A. Mohamed, K. Sopian, A review of solar energy modeling techniques, Renew. Sustain. Energy Rev. 16 (2012) 2864–2869.

[4] C. Voyant, M. Muselli, C. Paoli, M.L. Nivet, Optimization of an artificial neural network dedicated to the multivariate forecasting of daily global radiation, Energy 36 (1) (2011) 348–359.

[5] C. Voyant, G. Notton, S. Kalogirou, M.L. Nivet, C. Paoli, F. Motte, A. Fouilloy, Machine learning methods for solar radiation forecasting: a review, Renew. Energy 105 (2017) 569–582.

[6] S. Rehman, M. Mohandes, Artificial neural network estimation of global solar radiation using air temperature and relative humidity, Energy Policy 36 (2) (2008) 571–576.

[7] M. Bilgili, M. Ozoren, Daily total global solar radiation modeling from several meteorological data, Meteorology Atmos. Phys. 112 (3–4) (2011) 125–138.

[8] R. Belu, Artificial Intelligence Techniques for Solar Energy and Photovoltaic Applications, in: Handbook of Research on Solar Energy Systems and Technologies, IGI Global, 2013.

[9] A.K. Yadav, S.S. Chandel, Solar radiation prediction using Artificial Neural Network techniques: a review, Renew. Sustain. Energy Rev. 33 (2014) 772–781.

[10] A. Mellit, S.A. Kalogirou, Artificial intelligence techniques for photovoltaic applications: a review, Prog. Energy Combust. Sci. 34 (5) (2008) 574–632.

[11] A. Sozen, E. Arcakliogblu, M. Ozalp, Estimation of solar potential in Turkey by artificial neural networks using meteorological and geographical data, Energy Convers. Manag. 45 (2004) 3033–3052.

[12] M.A. Behrang, E. Assareh, A. Ghanbarzadeh, A.R. Noghrehabadi, The potential of different artificial neural network (ANN) techniques in daily global solar radiation modeling based on meteorological data, Sol. Energy 84 (8) (2010) 1468–1480.

[13] A.S. Dorvlo, J.A. Jervase, A. Al-Lawati, Solar radiation estimation using artificial neural networks, Appl. Energy 71 (4) (2002) 307–319.

[14] M. Bou-Rabee, S.A. Sulaiman, M.S. Saleh, S. Marafi, Using artificial neural networks to estimate solar radiation in Kuwait, Renew. Sustain. Energy Rev. 72 (2017) 434–438.

[15] M. Sahin, Y. Kaya, M. Uyar, S. Yidirim, Application of extreme learning machine for estimating solar radiation from satellite data, Int. J. Energy Res. 38 (2) (2014) 205–212.

[16] Y. Wu, J. Wang, A novel hybrid model based on artificial neural networks for solar radiation prediction, Renew. Energy 89 (2016) 268–284.

[17] H. Dong, L. Yang, S. Zhang, Y. Li, Improved prediction approach on solar irradiance of photovoltaic Power Station, TELKOMNIKA Indonesian J. Electr. Eng. 12 (3) (2014) 1720–1726.

[18] S. Salcedo-Sanz, C. Casanova-Mateo, A. Pastor-Sánchez, D. Gallo-Marazuela, A. Labajo-Salazar, A. Portilla-Figueras, Direct solar radiation prediction based on Soft-Computing algorithms including novel predictive atmospheric variables, Intelligent Data Eng. Automated Learn. - IDEAL 2013, Lect. Notes Comput. Sci. 8206 (2013) 318–325.

[19] K. Mohammadi, S. Shamshirband, C.W. Tong, M. Arif, D. Petkovic, S. Che, A new hybrid support vector machine-wavelet transform approach for estimation of horizontal global solar radiation, Energy Convers. Manag. 92 (2015) 162–171.

[20] L. Olatomiwa, S. Mekhilef, S. Shamshirband, K. Mohammadi, D. Petkovic, S. Ch, A support vector machine-firefly algorithm-based model for global solar radiation prediction, Sol. Energy 115 (2015) 632–644.

[21] J. Antonanzas, R. Urraca, F.J. Martinez-de-Pison, F. Antonanzas-Torres, Solar irradiation mapping with exogenous data from support vector regression machines estimations, Energy Convers. Manag. 100 (2015) 380–390.

[22] R.V. Monteiro, G.C. Guimarães, F.A. Moura, M.R. Albertini, M.K. Albertini, Estimating photovoltaic power generation: performance analysis of artificial neural networks, Support Vector Machine and Kalman filter, Electr. Power Syst. Res. 143 (2017) 643–656.

[23] J.L. Chen, H.B. Liu, W. Wu, D.T. Xie, Estimation of monthly solar radiation from measured temperatures using support vector machines - a case study, Renew. Energy 36 (2011) 413–420.

[24] J. Zeng, W. Qiao, Short-term solar power prediction using a support vector machine, Renew. Energy 52 (2013) 118–127.

[25] J. Zengand, W. Qiao, Short-term solar power prediction using a support vector machine, Renew. Energy 52 (2013) 118–127.

[26] S. Belaid, A. Mellit, Prediction of daily and mean monthly global solar radiation using support vector machine in an arid climate, Energy Convers. Manag. 118 (2016) 105–118.

[27] G. López, B. Batlles, J. Tovar-Pescador, Selection of input parameters to model direct solar irradiance by using artificial neural networks, Energy 30 (2005) 1675–1684.

[28] R. Yacef, M. Benghanem, A. Mellit, Prediction of daily global solar irradiation data using Bayesian neural network: a comparative study, Renew. Energy 48 (2012) 146–154.

[29] J. Wu, C.K. Chan, Y. Zhang, B.Y. Xiong, Q.H. Zhang, Prediction of solar radiation with genetic approach combing multi-model framework, Renew. Energy 66 (2014) 132–139.

[30] S. Salcedo-Sanz, C. Casanova-Mateo, J. Muñoz-Marí, G. Camps-Valls, Prediction of daily global solar irradiation using temporal Gaussian processes, IEEE Geoscience Remote Sens. Lett. 11 (no. 11) (2014).

[31] C.L. Fu, H.Y. Cheng, Predicting solar irradiance with all-sky image features via regression, Sol. Energy 97 (2013) 537–550.

[32] O. Senkal, T. Kuleli, Estimation of solar radiation over Turkey using artificial neural network and satellite data, Appl. Energy 86 (7–8) (2009) 1222–1228.

[33] R. Deo, M. Sahin, Forecasting long-term global solar radiation with an ANN algorithm coupled with satellite-derived (MODIS) land surface temperature (LST) for regional locations in Queensland, Renew. Sustain. Energy Rev. 72 (2017) 828–848.

[34] S. Bhardwaj, V. Sharma, S. Srivastava, O.S. Sastry, B. Bandyopadhyay, S.S. Chandel, J.R. Gupta, Estimation of solar radiation using a combination of Hidden Markov Model and generalized Fuzzy model, Sol. Energy 93 (2013) 43–54.

[35] M. Diagne, M. David, P. Lauret, J. Boland, N. Schmutz, Review of solar irradiance forecasting methods and a proposition for small-scale insular grids, Renew. Sustain. Energy Rev. 27 (2013) 65–76.

[36] M. Diagne, M. David, J. Boland, Post-processing of solar irradiance forecasts from WRF model at Reunion Island, Sol. Energy 105 (2014) 99–108.

[37] J. Wu, C.K. Chan, Prediction of hourly solar radiation using a novel hybrid model of ARMA and TDNN, Sol. Energy 85 (2011) 808–817.

[38] C. Voyant, M. Muselli, C. Paoli, M.L. Nivet, Hybrid methodology for hourly global radiation forecasting in Mediterranean area, Renew. Energy 53 (2013) 1–11.

[39] F.O. Hocaoglu, O.N. Gerek, M. Kurban, Hourly solar radiation forecasting using optimal coefficient 2-D linear filters and feed-forward neural networks, Sol. Energy 82 (2008) 714–726.

[40] F.J. Lima, F.R. Martins, Enio B. Pereira, E. Lorenz, D. Heinemann, Forecast for surface solar irradiance at the Brazilian Northeastern region using NWP model and artificial neural networks, Renew. Energy 87 (2016) 807–818.

[41] V. Sharma, D. Yang, W. Walsh, T. Reindl, Short term solar irradiance forecasting using a mixed wavelet neural network, Renew. Energy 90 (2016) 481–492.

[42] A. Aybar-Ruiz, S. Jiménez-Fernández, L. Cornejo-Bueno, C. Casanova-Mateo, J. Sanz-Justo, P. Salvador-González, S. Salcedo-Sanz, A novel grouping genetic algorithm - Extreme Learning Machine Approach for Global solar radiation prediction from numerical weather models inputs, Sol. Energy 132 (2016) 129–142.

[43] S. Salcedo-Sanz, C. Casanova-Mateo, A. Pastor-Sánchez, M. Sánchez-Girón, Daily global solar radiation prediction based on a hybrid Coral Reefs Optimization – Extreme Learning Machine approach, Sol. Energy 105 (2014) 91–98.

[44] J. Schmidli, C.M. Goodess, C. Frei, M.R. Haylock, Y. Hundecha, J. Ribalaygua, T. Schmith, Statistical and dynamical downscaling of precipitation: An evaluation and comparison of scenarios for the European Alps, J. Geophys. Res. 112 (2007) 1–20.

[45] S. Salcedo-Sanz, J. Muñoz-Bulnes, M. Vermeij, New Coral Reefs-based Approaches for the Model Type Selection Problem: A Novel Method to Predict a Nation's Future Energy Demand, Int. J. Bio-inspired Comput. (2016) in press.

[46] W.C. Skamarock, J.B. Klemp, J. Dudhia, D.O. Gill, D.M. Barker, W. Wang, J.G. Powers, A Description of the Advanced Research WRF Version 2, National Center for Atmospheric Reserach, Mesoscale and Microscale Meteorology Division, 2005. Technical Note.

[47] A.J. Litta, U.C. Mohanty, S.M. Idicula, The diagnosis of severe thunderstorms with high-resolution WRF model, J. Earth Syst. Sci. 121 (2) (2012) 297–316.

[48] D. Carvalho, A. Rocha, M. Gómez-Gesteira, C. Silva Santos, Sensitivity of the WRF model wind simulation and wind energy production estimates to planetary boundary layer parameterizations for onshore and offshore areas in the Iberian Peninsula, Appl. Energy 135 (2014) 234–246.

[49] S. Salcedo-Sanz, J. Del Ser, I. Landa-Torres, S. Gil-López, J.A. Portilla-Figueras, The Coral Reefs Optimization algorithm: a novel metaheuristic for efficiently solving optimization problems, Sci. World J. (2014) 739768.

[50] S. Salcedo-Sanz, A. Pastor-Sánchez, L. Prieto, A. Blanco-Aguilera, R. García-Herrera, Feature selection in wind speed prediction systems based on a hybrid coral reefs optimization–Extreme learning machine approach, Energy Convers. Manag. 87 (2014) 10–18.

[51] L. Cuadra, S. Salcedo-Sanz, J.C. Nieto-Borge, E. Alexandre, G. Rodríguez, Computational intelligence in wave energy: comprehensive review and case study, Renew. Sustain. Energy Rev. 58 (2016) 1223–1246.

[52] G.B. Huang, Q.Y. Zhu, C.K. Siew, Extreme Learning Machine: Theory and Applications, Neurocomputing 70 (1) (2006) 489–501.

[53] G.B. Huang, L. Chen, C.K. Siew, Universal approximation using incremental constructive feedforward networks with random hidden nodes, IEEE Trans. Neural Netw. 17 (4) (2006) 879–892.

[54] G.B. Huang, H. Zhou, X. Ding, R. Zhang, Extreme learning machine for regression and multi-class classification, IEEE Trans. Syst. Man Cybern. Part B 42 (2) (2012) 513–529.

[55] H.G. Beyer, J. Polo Martinez, M. Suri, et al., Deliverable 1.1.3. Report on Benchmarking of Radiation Products, Report under contract no. 038665 of MESoR, 2011, www.mesor.net/deliverables.html.

[56] J. Kleissl (Ed.), Solar Energy Forecasting and Resource Assessment, Academic Press, 2013.