ImPRoving aIr quality forecaSTs using advanced machIne learNing and big data tEchniques (PRISTINE)

Scientific proposal

Carlos Pérez García-Pando, *Roberto Serrano-Notivoli, Nicolau Manubens, Pierre-Antoine Bretonnière, Francesco Benincasa, Alasdair Hunter, Carles Tena, Javier Vegas, Laura Cifuentes, Alicia Sánchez Lorente, Kim Serradell and Maria Teresa Pay*

Barcelona Supercomputing Center

Index

1 Background . . . 2
1.1 Air quality forecasts for Europe and Spain . . . 2
1.2 Machine Learning and air quality forecasting . . . 5
2 Objectives . . . 6
3 Novelty, applicability and relevance . . . 6
4 Research methodology . . . 6
5 Experience and suitability of the research group . . . 10
5.1 Scientific expertise of the P.I. Dr. Carlos Pérez García-Pando . . . 11
6 Work plan and calendar . . . 12
7 Budget . . . 13
8 References . . . 13

Summary

Air quality forecasts typically rely on air quality models (AQMs) that simulate the lifecycle of air pollutants as they disperse and react in the atmosphere. Two critical aspects of these forecasts are their skill and their efficient provision before they lose their value. The skill of AQMs is typically limited by our incomplete knowledge of the relevant physical and chemical processes, the poor characterization of the initial conditions and the simplifications applied to reduce their computational burden. In this context, a smart combination of observations from air quality monitoring stations and model outputs has the potential of significantly improving the forecast skill. The Barcelona Supercomputing Center - Centro Nacional de Supercomputación (BSC-CNS) develops and operates the CALIOPE (CALIdad del aire Operacional Para España) air quality forecast system. CALIOPE is an integrated modelling framework to forecast air quality that accounts for pollutant emissions from the different activities and sectors, meteorology and the non-linear chemical transformations and interactions of atmospheric gases and particles. CALIOPE forecasts are corrected based on observations using Kalman filter techniques. Despite the positive effect of these techniques in the forecasts, they are far from perfect. In this context, the goal of PRISTINE is to enhance the CALIOPE air quality forecasts provided for Spain and Europe by optimally combining model outputs and observations based on computationally efficient machine learning and big data management techniques, including Clustering and Classification methods, and Neural Networks. The economic and social impact of the project is potentially large, since CALIOPE forecasts are a source of information widely used in Spain by mass media, companies, regional administrations, and the general public.

---

1 Background

According to the World Health Organization (WHO) ambient air pollution accounts for 3.5 million premature deaths per year globally, most of them happening in low- and middle-income countries. In Europe, emissions of many air pollutants have decreased over the past decades, resulting in improved air quality. However, exceedances of European regulatory limits of particulate matter, ground-level ozone and nitrogen dioxide are frequently registered, posing serious health risks for the population.

Forecasting air quality and understanding the causes of air pollution events requires the development and use of air quality models (AQMs). AQMs are complex mathematical representations used to simulate the physical and chemical processes affecting air pollutants as they disperse and react in the atmosphere based on meteorological data and pollutant emission inputs. AQMs are computationally demanding and require the use of supercomputers. **Daily air quality forecasts contribute to a better knowledge of the pollutants distribution and levels over populated areas and support decision-making and services directed towards reducing air pollution and/or mitigating its effects**. For example, forecasting accurately when and where the daily limit value for a certain air pollutant will be exceeded can allow public authorities to make upfront tailored decisions on traffic restrictions. On the other hand, the availability of these air quality forecasts through a website, a dedicated smartphone application, and/or social media (Facebook, Twitter, etc.), can allow thousands of people every day to decide, for example, whether they should reduce exercise at certain locations or times in order to mitigate the health risks from exposure to air pollution.

Two critical issues in air quality forecasting are skill and computational efficiency. It is well known that air quality forecast skills are limited by 1) our incomplete knowledge of relevant physical and chemical processes determining the fate of air pollutants, 2) errors in the input emissions and boundary conditions, and 3) a poor characterization of the model initial conditions. Skill is also partly related to computational efficiency: very often model schemes have to be simplified or the model resolution decreased to reduce the computational burden; otherwise the air quality forecasts could not be produced everyday in a timely manner.

In this context, our project **PRISTINE** aims at ImPRoving aIr quality forecaSTs using advanced machIne learNing and big data tEchniques. **Our goal is to enhance the widely used air quality forecasts produced by the CALIOPE system at the Barcelona Supercomputing Center - Centro Nacional de Supercomputación (BSC-CNS) by optimally combining model outputs and observations from thousands of air quality stations in Europe and Spain based on computationally efficient machine learning and big data management techniques**. We anticipate that these techniques have the potential to significantly correct the inherent errors in AQMs and provide improved air quality forecasts in a timely manner.

In this section we overview the CALIOPE air quality forecast system developed and operated at the BSC-CNS, we revise the current state-of-the-art, and we formulate the specific objectives of PRISTINE while highlighting its novelty, applicability, and relevance. The remaining sections describe the research methodology, the experience and suitability of the PI and the research group, and provide the work plan, schedule and budget.

1.1 Air quality forecasts for Europe and Spain

The BSC-CNS develops and operates an integrated modelling framework to diagnose and forecast air quality that accounts for the emissions from the different activities and sectors, meteorology and the non-linear chemical transformations and interactions of atmospheric gases and particles. This modelling framework, so-called CALIOPE system (CALIdad del aire Operacional Para España), runs at high spatial and temporal resolution in the MareNostum IV supercomputer. CALIOPE is an operational system (www.bsc.es/caliope) providing air quality forecasts every day over Europe and Spain.

The CALIOPE air quality system (Figure 1) provides 24- and 48-h forecasts for regulated pollutants (ozone, nitrogen dioxide, sulfur dioxide, and particulate matter) with high spatial resolution over Europe,

the Iberian Peninsula and Canary Islands. It consists in a set of models: the WRF-ARW meteorological model; the HERMES emission model, the BSC-DREAM8b natural dust model, and the CMAQ. Europe is solved at 12 km spatial resolution and the Iberian Peninsula is solved at higher spatial resolution (4km) (Pay et al., 2014; Schaap et al., 2015).

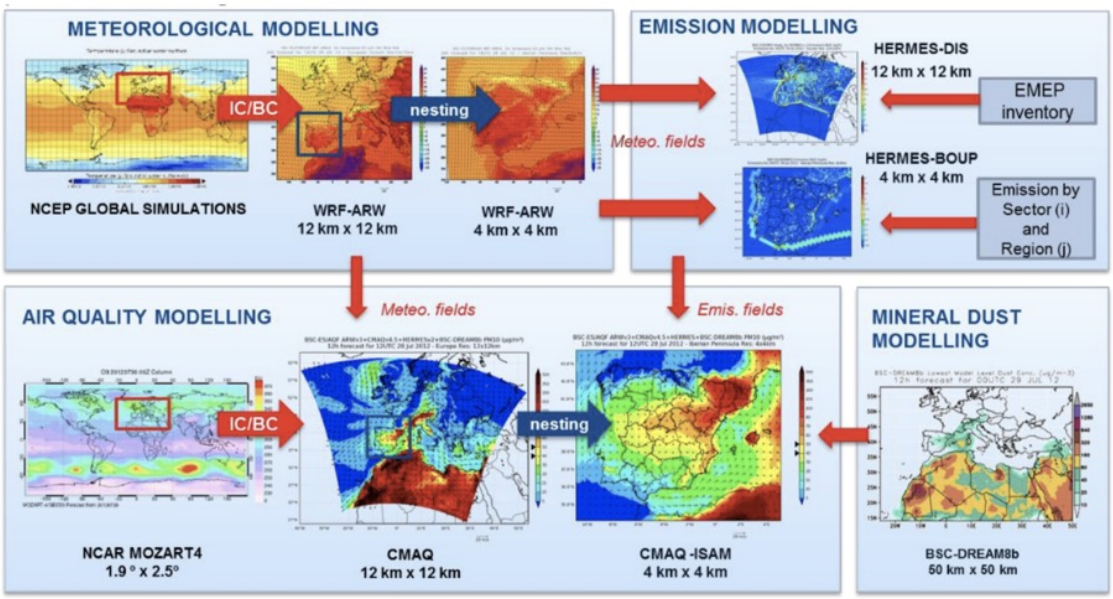

Fig. 1: CALIOPE modeling framework

Each of the three components of the model represent the key dependencies of the air quality:

1. The WRF-ARW **meteorological** model is widely used for multiple purposes such as weather (Davis et al., 2008; Skamaroc et al., 2008) and air quality (Pay et al., 2010b, Baldasano et al., 2011; Gonalves et al., 2009a) and it is used downscaled to each target spatial extent with the aim of be representative of the regional atmospheric dynamics.

2. The HERMES (High-elective Resolution Modelling Emission System) **emission** model (Baldasano et al., 2008b) generates the emissions for Spain needed for the application of high-resolution chemistry transport models and it is capable of calculating emissions by sector-specific sources or by individual installations and stacks.

3. The system is based on the **air quality model** CMAQ, which has been widely evaluated during its development over northeastern Spain (Jiménez et al., 2005a,b, 2006a,b, 2007), the Iberian Peninsula (Jiménez-Guerrero et al., 2008; Baldasano et al., 2008a, 2011; Pay et al., 2010a, 2012a) and Europe (Pay et al., 2010b; Basart et al., 2012; Pay et al., 2012b). Furthermore, it has been used for assessing on the contribution of atmospheric processes affecting the dynamic of air pollution (Gonalves et al., 2009a) and as management tool to study air quality impact of urban management strategies (Jiménez-Guerrero et al., 2008; Gonalves et al., 2008; 2009b; Soret et al., 2011, 2013, 2014; Baldasano et al., 2014). CALIOPE also has been recently used to assess air pollution effects on health in Spain (Aguilera et al., 2013; Akita et al., 2014).

The CALIOPE evaluation studies demonstrated that the system, which does not only rely on the meteorological prediction but also on chemical transport modelling and on highly uncertain emission inventories, is likely to have significant model errors.In order to improve air quality forecast skills before dissemination to the public, CALIOPE uses a Kalman filter (Sicardi et al., 2012) method to reduce systematic errors in air quality predictions. This post processing technique combines the model outputs with observations provided in near-real time by the European Environment Information and Observation Network (EIONET; https://www.eionet.europa.eu/) to correct the forecast. EIONET provides a dense geographical coverage of Europe and the Spanish territory (Figure 2).

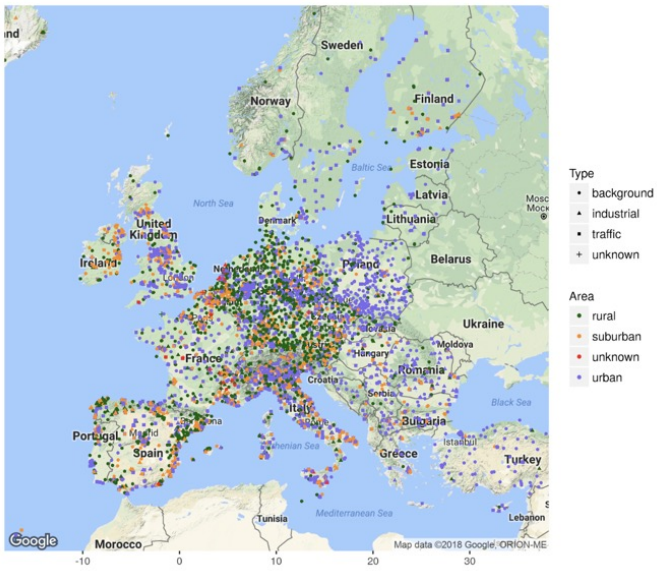

Fig. 2: EIONET (European Environment Information and Observation Network) stations network over Europe measuring air quality. Stations are differenced by pollutant origin (shapes) and type of location environment (colours).

Consequently, the forecast method can be summarized in three stages: 1) A first model output based on the atmospheric dynamics (WRF-ARW), the emissions (HERMES) and the chemistry (CMAQ) is produced; 2) a post-processing process of bias correction of the model based on the observations through a Kalman filter is performed and; 3) a new model output is produced based on this correction, which provides a final air quality forecast with a real-time correction to adapt the predictions to the reality (Figure 3).

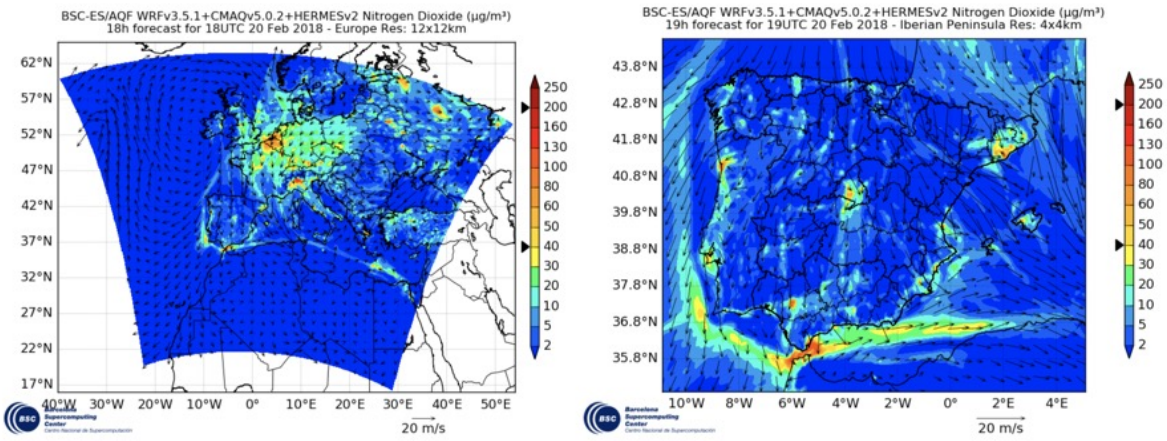

Fig. 3: 48-h forecast from the CALIOPE system for surface hourly nitrogen dioxide (NO2) concentrations (in μgm-3) at 18UTC on February 20th 2018 in Europe (left) and in the Iberian Peninsula (right). Exposure to NO2 can cause respiratory effects in healthy people and increased respiratory symptoms in people with asthma. NO2 creates ozone (O3) which exacerbates respiratory conditions and other ailments.

1.2 Machine Learning and air quality forecasting

In its current form, CALIOPE model handles the post-processing process of bias correction based on a Kalman filter (KF) (Sicardi et al., 2012), which corrects of the daily model forecasts using the point-based observations all over the spatial extent (Figure 2). The KF corrects the systematic under/overestimations of the model, and is capable of adapting parameters to new situations, unless drastic changes are occurring over the time. The correction of this bias is applied to the locations with observations using the hourly values of the model over the previous 15 days and the corresponding observations. The KF is a linear, adaptive, recursive and optimal algorithm; it works by a mechanism of prediction and correction of the bias (between the model results and the observations) at each time step. This method greatly improves the forecasts of AQ (Figure 4) thanks to the use of real-time observations.

This post-processing technique is useful but it is far from perfect. As currently implemented, the KF is able to adapt its coefficients as new forecasts and observations become available. However, it unables to correct sudden extremely biased events, i.e., to anticipate a large forecast error when all errors during the past few days have been smaller. This happens for example when a strong anticyclone affects a specific region and the air quality model does not correctly capture the low dispersive conditions. This particular situation has prompted the need for advanced and dynamic methods such a Machine Learning techniques to better predict sudden peaks in air pollution. **In PRISTINE we propose to use of Clustering and Classification methods, as well as Neural Networks, which have been proved to be useful in other scientific areas, to improve the air quality forecasts by combining forecast model outputs and observations**.

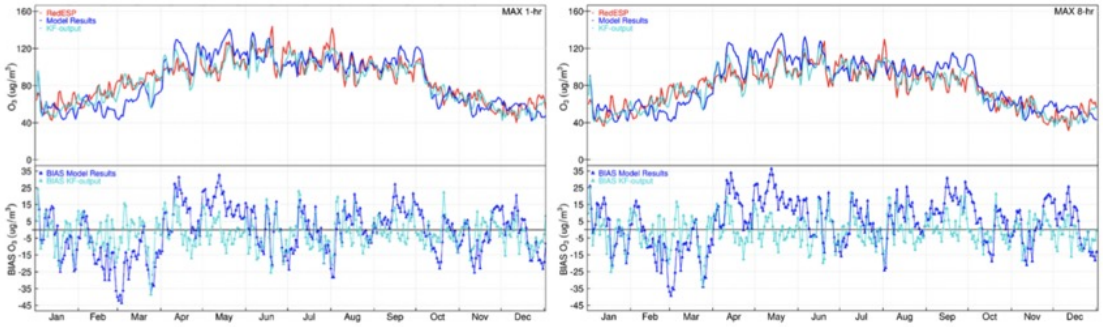

Fig. 4: Time series of the max. 1-h (left panel) and max. 8-h (right panel) ground level O3 concentration (μg m-3), averaged for all the Spanish network of stations (RedESP) for the model results (CALIOPE raw forecast) and the Kalman filter (CALIOPE corrected forecast), for the year 2004. The plot of the biases (μg m-3) is also included.

In addition to improving air quality forecast skills, a critical factor is computational efficiency. Real-time services are steadily providing great volumes of information. In our case the post processing to improve forecasts needs a variety of inputs, from observations collected from thousands of monitoring stations to input from the meteorological and air quality models at different spatial and temporal resolutions. The huge size and variety of the input datasets presents a challenge, especially in a real-time operational context. For instance, the whole system needs enough time to pre-process, calculate, post-process, store and disseminate daily forecasts before dealing with the next day. The critical aspect is that predictions have to be available as soon as possible so that they can be useful in relevant decisions before they lose their value. **In PRISTINE we will explore different types of technologies including the use workflow tools, combinations of fat and thin nodes, and a clever mix of disk and tape and HPC (High Performance Computing) resources to make our developed post processing algorithms and data management more efficient.**

2 Objectives

The main objective of PRISTINE is **to use Machine Learning and Big Data techniques to improve air quality forecasts over Europe and Spain**. To achieve the main goal, the work will proceed in three stages:

(1) We will first explore machine learning techniques such as ***Clustering and Classification*** **methods and** ***Neural Networks*** to improve the Kalman Filter post-processing that combines air quality model outputs and observations.

(2) In a second step, we will **test, evaluate and optimize our new methods** for the European and Iberian Peninsula forecasts domains, especially **focusing on urban areas**.

(3) Finally, we will **implement the improved post processing in the CALIOPE operational suite**.

3 Novelty, applicability and relevance

PRISTINE will enhance air quality (AQ) forecasting in the new era of the Big Data. The use of Machine Learning Techniques is highly innovative in a field that has traditionally devoted most of its effort to improve the representation of the physical and chemical processes in air quality models. Machine learning can indeed play an important role in the enhancement of air quality forecasts, since it helps identifying patterns, and consequently allowing the extraction of valuable information from the data with unprecedented accuracy. Also Big Data management for these applications can benefit from advanced analytics techniques inspired by machine learning methods. In this sense, PRISTINE will take advantage of novel data retrieval methods and HPC architectures to achieve the best outcomes when dealing with large amount of datasets in almost-real time.

In addition to these technical contributions and novelties, the economic and social impact of the project is potentially large. The most clear benefit will be the direct improvement of the real-time forecasts of AQ that will enhance the alerts to population, especially in those urban areas more vulnerable to these kind of events: the high populated cities. We note that the CALIOPE air quality forecast system is widely used in Spain and provides support to regional administrations in their decision-making processes. The air quality information shown in the website (Figure 3) is used by Spanish TV for meteorological news (e.g., TVE1, la Sexta, Antena 3) and companies (e.g. refineries, cement industry sector, power generation companies). The forecasts are freely available on the web-portal wwww.bsc.es/caliope (45.185 visits in 2017) and through the CALIOPE smartphone application (15.816 users). The vast majority of the users are from Spain (90.67%), especially from the cities of Madrid and Barcelona (Google analytics). We also disseminate relevant information on air quality that the Spanish citizen will breathe, via the recently created CALIOPE twitter account (@BSC_CALIOPE).

4 Research methodology

We propose a novel post-processing procedure to improve the performance of the KF discussed in previous sections. We seek methods able to adapt, even with a short training periode, to abrupt changes in weather conditions or sudden changes of forecast error especially when rapid transitions of biased events are likely to occur. One of the new methods will rely on the analog concept. The analog of a forecast for a given location and time is defined here as a past prediction that matches selected features of the current forecast.

Analogs will be searched across multiple physical variables and time windows for a given location and forecast time. Good analogs are considered as forecasts that predicted similar values and temporal trends of the forecasted quantity for which the error needs to be estimated (e.g pollutants concentration in this study), and for other variables that exhibit correlation to the latter (e.g, pressure, temperature,

humidity, etc). We will assume that if forecasts in the past can be found that are similar to the current prediction, then the current prediction error can be inferred by analyzing the errors of the analogs. In contrast to previous studies (e.g. Djalalova et al., 2015), which considered standard statistics to establish when a good analog forecast was found, **we will use machine learning algorithms like clustering methods for pattern recognition within the forecast datasets and robust classification methods like random forest and deep learning** to perform dynamic assignments accordingly and time-series predictions, which will benefit from the hybrid Marenostrum 4 architecture (e.g GPUs for code acceleration).

The work is organized around two technical work packages (WP) and a project management and dissemination work package as described below.

WP1. Establishment of an optimal analog methodology based on machine learning techniques

The method developed through WP1 should overcome the disadvantages of the current implementation of the KF. In order to predict large day-to-day changes in the prediction error we will run the KF through an ordered set of analogs rather than through the previous 15 days. Analogs here are defined as past forecasts that are similar to the current forecast (as measured by a particular metric) for which a correction in real time is desired. To have an ordered set of analogs means that the analogs are ranked (from left to right) from the farthest (worst analog) to the closest (best analog). By running the KF through the ordered analogs the correction for the current forecast will give more weight to the analog forecasts closer to it, resulting in a better correction. Therefore defining and finding proper analogs is really crucial.

Another method we will test is the use of recurrent and bidirectional networks for forecasting time-series. Recurrent neuronal networks are a powerful type of neural network designed to handle sequence dependence, in this case time series, and in addition are able to almost seamlessly model problems with multiple input variables. Instead of the Low-Short-Time-Memory (LSTM) layer used in previous works (Zhang, 2017), we will use the Gated Recurrent Unit (GRU) layer developed by Chung et al. 2014. Gated recurrent unit (GRU) layers work using the same principle as LSTM, but is more efficient from the computational point of view. The technique that will be utilized here for the search of patterns, as well as for the time-series prediction, is based on the combination or ensemble of two recurrent neuronal networks; one working chronologically along the time-series and the second running reversely. The **bidirectional neuronal network (BNN)** will be trained to capture the temporal relationship among time series data, and after validation, will deliver a model capable for prediction. For this purpose it is also equally important to determine how long the historical observations are to be used to predict the future. Of course the longer the length is, the better the prediction will be but also more computationally expensive. This apparent problematic issue will be efficiently managed by using HPC resources. In analogy with the analogs-method, the design of the metrics is also crucial, for what parameters such as number of fully-connected layers and hidden units plays a significant role. The results extracted by the analog-method could be used as baseline or quality check for this particular technique.

Task 1.1. Determination of suitable metrics for definition and discovery of forecast analogs and time series forecasting

The metric which will be defined here should account for similarities in the errors between the current forecast to be corrected at a given time and station location, and the past predictions or analogs, at times before the current forecast was issued, and the same location. The analogs will be created using **clustering algorithms** like k-nearest neighbors or k-means (Turco et al. 2017) which rely on the euclidean distances between their members. We will first explore the use of analogs based on the errors of the pollutant concentrations in previous forecasts cycles. Once the analogs are established, a **classification random forest** will be applied to both, the current forecast and the generated clusters, to find out which are the most valuable or closest analogs among the generated clusters. By using a random forest with regression, a simple metric like the Manhattan distance metric ($|A_i| - |A_j|$) could be applied in order to perform the assignment accordingly. Once we have the analogs, the KF will be applied to the set of ordered analogs in order to correct the model output.

For the case of **time-series forecasting**, the previously designed neuronal structure will be optimized (trained and validated) according to the number of fully-connected layers and hidden units. Aspects like learning rate and outfitting are going to be extensively evaluated here in order to deliver the best model performance by using the current forecast and the root mean squared error (RMSE).

Task 1.2. Sensitivity analysis for the analog-based method

In addition to using the errors in the concentration of pollutants for the analog determination, we will evaluate to what extent similar meteorological scenarios can be used as analogs, and particularly how the skill of the forecast depends on the variables included in the analogs (temperature, solar radiation, wind speed and direction, etc). To that end, multivariate clustering techniques will be applied, which is expected to reduce the ambiguities due to the use of only one variable for the analogs search, resulting in improvements of the forecast accuracy.

Task 1.3. Metrics for the evaluation of the performance of the post-processing method

For a first evaluation of these methodologies, hourly-datasets generated in the past two years by the CALIOPE model and observed data from a set of representative stations in Spain will be used. The forecast variable to be corrected will be for instance ozone, and can be extended to the rest of forecast variables such as nitrogen dioxide, sulfur dioxide, and particulate matter. In order to evaluate the performance of the different methods a set of evaluation metrics will be applied, such as the centered root-mean square error (CRMSE), the BIAS for systematic error evaluation, correlation and normalized standard deviation (NSD) to generate Taylor's diagrams (Taylor 2001) and Rank correlation for measuring the strength of monotone associations between two variables (prediction and observation).

WP2. Metrics evaluation as a function of time and space and real-time forecasting

Once we have tested a variety of analog and BNN methods using a reduced set of stations in Spain, we will select the method or methods showing the best performances. We will then compute and evaluate them using all the stations at national and European scale. We will also interpolate the station corrected model forecasts to the entire model grid, allowing for display of 2D post-processed maps of pollutants and aerosols (Kriging).

The main challenge here will be to extend our evaluation metrics from national to the european scale, since that would incur in a huge amount of data, which has to be treated accordingly and effectively to provide proper real-time air quality forecast. Dealing with multi-dimensional data sets (commonly used in the earth sciences community) requires to have an optimum working strategy to extract the best outcomes at the scientific and at the computational level. As a consequence, highly efficient flow diagrams have to be designed and implemented ahead in order to benefit from the computational advantages of working with HPC architectures. For instance, strategies based on the dynamical splitting of raw data (chunking) will be utilized to take advantage of distributed computer arquitectures. Our team in BSC-ES has already worked in this research line through different approaches using R language (*startR* package), providing a new insight on Big Data retrieval. Also, some other solutions have been proposed using Python (*xarray*), but they need to be tested to check if they fit with the Big Data used in our use case.

Task 2.1. Search for solutions to load and handle large data multivariate files from different sources, type of variables and number of dimensions in HPCs.

Different kind of scenarios will be addressed here in order to evaluate the memory and time consumptions of this particular task, in particular at regional, national and european scales. The evaluation will be done making use of the Marenostrum IV ressources, and taking benefit from extended existing chunking functions, allowing a dynamical treatment of the multi-dimensional datasets. In addition, the performance of the post-processing workflow will be accordingly validated and optimized by utilizing different tools in both, Python and R frameworks. Visualization tools like MapGenerator, will also be upgraded in order to enable

the visualization of both, pollutants and aerosols concentration, and meteorological variables in an interactive way.

Task 2.2. Display of 2D-post-processed maps by improving the Kriging method making use of the data processing and visualization techniques of task 2.1

As mentioned above, in order to display a 2D-post-processed map, accurate interpolation has to be made according to the corrected forecast in all the stations. For this specific topic, the performance of the KF + Analogs will be evaluated not only as a function of time, but also as a function of space. That means for instance, an analysis of the spatial distribution of the methods skill here presented, where the metrics are computed with all available pairs of observations and predictions at each station and for all stations available (e.g starting from regional to european scale). The robustness of the new analog-based method will then be evaluated and tested across a range of conditions, regardless of the different topographical and land-use characteristics of the locations considered.

Task 2.3 Improving the CALIOPE forecast service

The post processing developed will be implemented in the CALIOPE forecast system (Figure 5). The best identified techniques will be applied to improve the forecasts of the air quality at high spatial resolution over Europe (12 km) and Spain (4 km). The operational test of the online Operational Service will be carried out by making use of independent datasets in order to test the performance in terms of timing, memory consumption level and forecast accuracy. Figure 5 represents the current post-processing workflow for Spain. The new post-processing workflow will additionally include the new methods developed to combine model outputs and observations. The outcomes of the entire post-processing will result in the interactive visualization of timeseries and the display of 2D-maps for pollutants concentrations and meteorological variables.

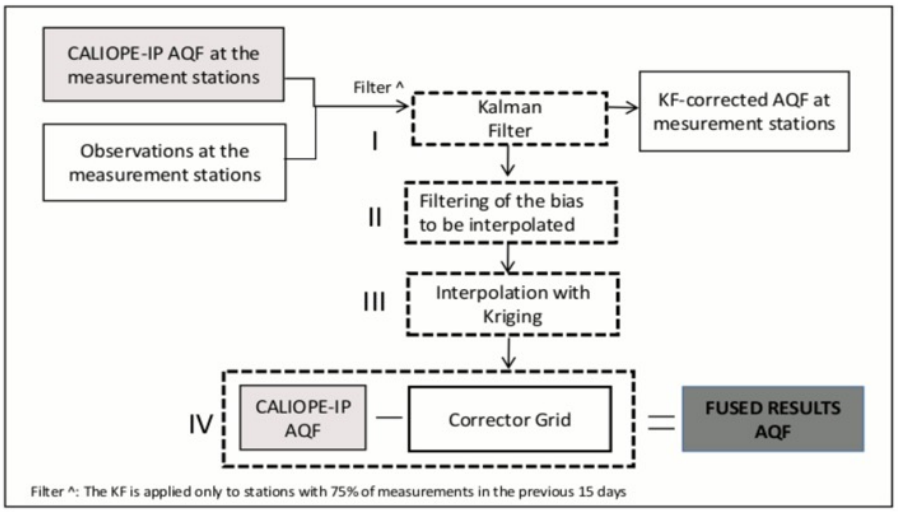

Fig. 5: Current Post processing workflow. Steps followed to adjust the air concentration fields from the CALIOPE system. IP stands for Iberian Peninsula, AQF stands for Air Quality Forecasts

WP3. Project management and dissemination of results

This work-package will ensure the appropriate management of the project and broadly disseminate the outputs throughout its duration. It will be feasible thanks to administrative support at the host institution (BSC) and the strong team Earth Science Services established at departmental level (BSC-ES). WP4 will monitor the progress of the project, ensure timely preparation of scientific reports and

outreach activities, facilitate communication among the Research Team members and organize the project meetings. The project PI has a large experience in project management; he has managed large research projects funded by NASA, NOAA and the Department of Energy in the US.

Task 3.1. Scientific dissemination.

The results of the proposed work will be disseminated through several channels. First, results will be prepared as a series of scientific articles submitted for publication in journals such as *Environmental Modelling & Software*, *Computers & Geosciences* or *Journal of Advances in Modeling Earth Systems*. The research group has a history of publishing in these and other high impact journals; Earth Sciences Department of BSC has produced more than 100 scientific publications in the last three years. Results from the proposed work will be presented at EGU, AGU and other appropriate international conferences.

Task 3.2. General public communication.

The communication of research results and public engagement will be a key focus of the proposed work, and it is included in our project. The research group has a strong history in presentation of research results at scientific conferences, and engagement with the public through publications to the media through the BSC Communication department. Indeed, BSC has dedicated staff and several operational programs in place to communicate activities to other researchers, students, and the general public that will be exploited by the research group as part of the project. The BSC operates as a PRACE Advanced Training Centre with a mission to provide training and education related to utilization of European supercomputing resources, including for environmental simulation. As part of the project, members of the research group will participate in the PRACE training program modules related to atmospheric modelling. Finally, results will be presented approximately once a year as part of the BSC Research Seminar Lecture series.

5 Experience and suitability of the research group

The **Earth Sciences department** of the BSC-CNS (BSC-ES) conducts multi-faceted research in Earth system modelling. It is structured around four groups with more than 65 members that have published more than 150 research articles in peer-reviewed journals over the last 5 years and with a very dense international collaborative network counting at least 50 institutes worldwide. BSC-ES focuses research on atmospheric emissions, air quality, mineral dust transport, computational efficiency of air quality and climate codes, data storage, analysis and dissemination, and global and regional climate modelling and prediction. The BSC-ES works on the development of and conducts research with a multi-scale set of comprehensive single-component and coupled regional and global models. During last 5 years (2013-2017), BSC-ES was granted 9 EU H2020 projects, 5 EU FP7 projects, 5 EU Copernicus projects, 10 projects funded by the Ministerio de Economía y Competitividad (MINECO), 2 projects funded by the European Space Agency, 1 project funded by the French Ministry of Sciences and 1 project from ERA-NET.

The BSC-ES international activity includes the coordination of the two World Meteorological Organisation (WMO) regional centres specialised in sand and dust warning and forecasting, as well as the participation in climate services initiatives like the Climate Services Partnership (CSP). Members of the BSC-ES participate in committees of the World Climate Research Programme (WCRP), such as the CLIVAR Scientific Steering Group or the Working Group on Seasonal to Interannual Prediction (WGSIP).

PRISTINE is a collaboration between the Atmospheric Composition (AC) group, led by the principal investigator Carlos Pérez García-Pando and the Computational Earth Sciences (CES) group, where most of the research team members belong.

In PRISTINE, the AC group provides both the expertise in air quality forecasting, acquired over the last 10 years of research, and the identification of the challenges in the post-processing correction of

the CALIOPE model. The CES group contributes with the experience of working in High Performance Computing (HPC) architectures using advanced Descriptive Analytics of large data volumes and novel explorative methods including deep learning and cognitive algorithms for pattern recognition in large and variate data sets. The work proposed aims to find an optimal solution for the provision of reliable air quality forecasts, taking advantage of the unique position of BSC, which is both an High Performance Computing (HPC) facility and an Earth Sciences service provider. A new research line in this context has been created in last year with the aim of further developing tools to operate in HPC architectures. A summary of the existing tools and their active research line are shown in Table 1.

Name	PL	C	RL	S
s2dverification	R	Climate models verification	1, 2, 3, 4, 5	https://goo.gl/WVKwtx
mapGenerator	Python	2D plotting for Earth Sciences datasets	1	https://goo.gl/wZQEsn
startR	R	Transform and arrange multidimensional data sets	2, 3	https://goo.gl/2QKu37
multiApply	R	Applies calculations over multidimensional arrays	2	https://goo.gl/bDRKBjx
easyNCDF	R	Simplifies NetCDF writing/reading processes	1	https://goo.gl/g4nBsB
ET - Evaluation Tool	R	Evaluation of atm. comp. simulations	3, 4	https://goo.gl/wLGS3J

Tab. 1: Tools developed by BSC for climate/air quality verification and analysis. PL: Programming Language; C: Capabilities; RL: Research Lines; S: Source. The research lines are all based in the use of large fat nodes: 1: Integrate the loading of Big Data files; 2: Process large files by splitting them in chunks; 3: Operate with unlimited functions in a simple way; 4: Add complex statistical analysis for models verification; 5: Integrate machine learning methods for climate analysis.

A great advantage for the project is the immediate availability of use of the Marenostrum IV supercomputer. It has a performance capacity of 13.7 Petaflop/s and is composed of two distinct parts. The general-purpose element, provided by Lenovo, has 48 racks with more than 3,400 nodes with next generation Intel Xeon processors and a central memory of 390 Terabytes. Its peak power is over 11 Petaflop/s, i.e. it is able to perform more than 11,000 trillion operations per second, ten times more than MareNostrum III despite costing only a 30% more in energy consumption. The second element of MareNostrum IV is formed of clusters of three different technologies that will be added and updated as they become available. These are technologies currently being developed in the USA and Japan to accelerate the arrival of the new generation of pre-exascale supercomputers. MareNostrum IV will have a disk storage capacity exceeding 10 Petabytes and will be connected to the Big Data infrastructures of BSC, which have a total capacity of 24.6 Petabytes.

5.1 Scientific expertise of the P.I. Dr. Carlos Pérez García-Pando

The principal investigator of the project is Dr. Carlos Pérez García-Pando. After 8 years as researcher at the NASA Goddard Institute for Space Studies and Columbia University, Dr. Perez joined the BSC in October 2016 as Head of the Atmospheric Composition group and AXA Professor. His research focuses on understanding the physical and chemical processes controlling atmospheric aerosols and gases, and on evaluating their effects upon climate, ocean biogeochemistry, air quality and health. He is also a model developer with large experience in supercomputers and operational air quality forecasting. He is considered one of the internationally leading figures in dust aerosol science. Thanks to the international recognition of his contributions to basic, applied and cross-disciplinary aspects within his field, he was awarded with an **AXA Chair (Cátedra AXA)** to develop his research program at the Barcelona Supercomputing Center. He also obtained a national **Ramón y Cajal position**, in which he was **ranked first by the Earth Sciences panel**. He was recently granted with an **ERC Consolidator Grant** that will start in October 2018. It is the only ERC Consolidator granted to a Spanish researcher in 2017. Also in 2017 the Royal Academy of Engineering (RAI, in its Spanish acronym) awarded Carlos Pérez García-Pando with the **Agustín de Betancourt y Molina prize** for young researchers for his contributions in the field of environmental risks.

6 Work plan and calendar

The table below shows the chronogram of the project over its 24 months' expected duration. Deliverables (D) and Milestones (M) are indicated and described below, and the contribution of each member of the group is detailed in the different tasks.

Research members team: *Carlos Pérez García-Pando (CP) (Project coordinator), Roberto Serrano (RS), Nicolau Manubens (NM), Pierre-Antoine Bretonnière (PB), Francesco Benincasa (FB), Alasdair Hunter (AH), Carles Tena (CT), Javier Vegas (JV), Laura Cifuentes (LC), Alicia Sánchez Lorente (AS), Kim Serradell (KS), M Teresa Pay (MP) and postdoc (PD)*

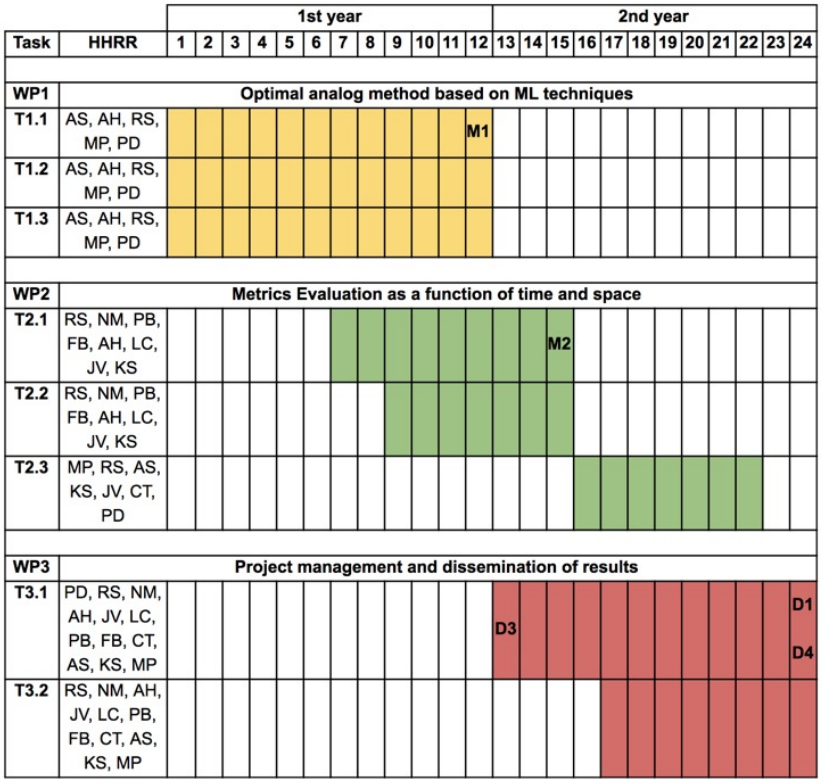

Fig. 6: Chronogram of the project

Milestones and Deliverables

Milestones:

M.1. Completion of an analog method based on ML techniques.

M.2. Evaluation of the KF + analogs method for all the situations

M.3. Application of the developed method to the CALIOPE operational service

Deliverables:

D.1. Description of the KF + analog post-processing method in AQ forecasts (scientific publication).

D.2. Benchmarking of the current AQ forecasts and the improved with the new method (technical report and conference communication).

D.3. First year report summarizing the advances.

D.4. Final report summarizing the scientific achievements of the project.

7 Budget

We request budget for hiring a Research Engineer (postdoc data scientist) during the first 20 months of the project. The hired postdoc will mainly focus on WP1 and will support other development tasks in WP2. The dissemination and communication tasks of WP3 will be also carried out in part by the hired postdoc (focusing on the scientific publication and the conferences). Two open-access publications are expected to show the main results of the project. We will also participate at least at two EGU General Assembly conferences (2019 and 2020).

Personnel costs	**Description**	**Cost**	**Justification**
	Research Engineer position	63,300.00 €	1 Research Engineer (20 months)
Other costs	**Description**	**Cost**	**Justification**
	Abstracts and posters	90.00 €	Submissions + material
	Conference fees	690.00 €	Fees: 2 EGUGeneral Assembly
	2 Open-Access publications	2,000.00 €	
Travel	**Description**	**Cost**	**Justification**
	Travels	2,000.00 €	2 EGU General Assembly
Direct costs		**68,080.00 €**	
Indirect costs		**31,820.91 €**	
Total		**99,900.91 €**	

Tab. 2: Budget

8 References

Aguilera, I., Basagaña, X., Pay, M.T., Agis, D., Bouso, L., Foraster, M., Rivera, M., Baldasano, J.M., Kunzli, N. 2013. Evaluation of the CALIOPE air quality forecasting system for epidemiological research: The example of NO2 in the province of Girona (Spain). *Atmospheric Environment*, 72: 134-141. DOI: 10.1016/j.atmosenv.2013.02.035

Akita, Y., Baldasano, J.M., Beelen, R., Cirach, M., de Hoogh, K., Hoek, G., Nieuwenhuijsen, M., Serre, M.L., de Nazelle, A. 2014. Large Scale Air Pollution Estimation Method Combining Land Use Regression and Chemical Transport Modeling in a Geostatistical Framework. *Environmental Science & Technology*, 48 (8): 4452-4459. DOI: 10.1021/es405390e

Basart, S., Pay, M. T., Jorba, O., Pérez, C., Jiménez-Guerrero, P., Schulz, M., and Baldasano, J. M. 2012. Aerosols in the CALIOPE air quality modelling system: evaluation and analysis of PM levels, optical depths and chemical composition over Europe, *Atmos. Chem. Phys.*, 12: 3363-3392. DOI:10.5194/acp-12-3363-2012

Baldasano, J.M.,Paya, M.T., Jorba, O., Gassó, J., Jiménez-Guerrero, P. 2011. An annual assessment of air quality with the CALIOPE modeling system over Spain. *Science of The Total Environment*, 409: 2163-2178. DOI: 10.1016/j.scitotenv.2011.01.041

Baldasano J.M, Jiménez-Guerrero, P., Jorba, O., Pérez, C., López, E., Gereca, P., Martin, F. , García-Vivanco, M., Palomino, I., Querol, X., Pandolfi, M., Sanz, M.J., Diéguez, J.J. 2008a. CALIOPE: An operational air quality forecasting system for the Iberian Peninsula, Balearic Islands and Canary Islands-First annual evaluation and ongoing developments. *Advances in Science and Research*, 2: 89-98, ISSN: 1992-0628.

Baldasano J.M., L. P. Gereca, E. López, S. Gassó, P. Jimenez-Guerrero. 2008b. Development of a high-resolution (1 km x 1 km, 1 h) emission model for Spain: the High-Elective Resolution Modelling Emission System (HERMES). *Atmospheric Environment*, 42 (31): 7215-7233, DOI:10.1016/j.atmosenv.2008.07.026

Chung, J., Gulcehre, C., Cho, K., Bengio, Y. 2014. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling, arXiv:1412.3555.

Djalalova, I., Delle Monache, L., Wilczak, J.2015. PM2.5 analog forecast and Kalman filter post-processing for the Community Multiscale Air Quality (CMAQ) model. *Atmospheric Environment*, 108: 76-87. DOI: 10.1016/j.atmosenv.2015.02.021

Davis, C., W. Wang, S.S. Chen, Y. Chen, K. Corbosiero, M. DeMaria, J. Dudhia, G. Holland, J. Klemp, J. Michalakes, H. Reeves, R. Rotunno, C. Snyder, and Q. Xiao, 2008: Prediction of Landfalling Hurricanes with the Advanced Hurricane WRF Model. *Mon. Wea. Rev.*, 136, 19902005,DOI: 10.1175/2007MWR2085.1

Goncalves, M., Jimenez Guerrero, P., Lopez, E., Baldasano, J.M., 2008. Air quality models sensitivity to on road traffic speed representation: Effects on air quality of 80 km h 1 speed limit in the Barcelona Metropolitan Area. *Atmospheric Environment*, 42, 8389 8402.

Gonalves, M., P. Jiménez-Guerrero, and J. M. Baldasano. 2009a. Contribution of atmospheric processes affecting the dynamics of air pollution in South-Western Europe during a typical summertime photochemical episode. *Atmospheric Chemistry and Physics*, 9, 849-864, www.atmos-chem-phys.net/9/849/2009/.

Goncalves, M., Jimenez Guerrero, P., Baldasano, J.M., 2009b. High resolution modeling of the effects of alternative fuels use on urban air quality: Introduction of natural gas vehicles in Barcelona and Madrid greater areas (Spain). *Science of the Total Environment* 407, 776 790.

Jiménez, P., Parra, R., Gassó, S., Baldasano, J.M. 2005b. Modeling the ozone weekend effect in very complex terrains: a case study in the Northeastern Iberian Peninsula, *Atmospheric Environment*,39 (3): 429-444. DOI: 10.1016/j.atmosenv.2004.09.065

Jiménez P, Jorba O, Parra R, Baldasano JM. 2005a. Influence of high-model grid resolution on photochemical modelling in very complex terrains. *Int J Environ Pollut*, 24: 180200.

Jiménez, P., Jorba, O., Parra, R., Baldasano, J.M. 2006a. Evaluation of MM5-EMICAT2000-CMAQ performance and sensitivity in complex terrain: High-resolution application to the northeastern Iberian Peninsula. *Atmospheric Environment*, 40 (26): 5056-5072. DOI: 10.1016/j.atmosenv.2005.12.060

Jiménez P, Lelieveld J, Baldasano JM. 2006b. Multiscale modeling of air pollutants dynamics in the northwestern Mediterranean basin during a typical summertime episode. *J Geophys Res*, 111: D18306. DOI:10.1029/2005JD006516.

Jiménez, P., Parra, R., Baldasano, J.M. 2007. Influence of initial and boundary conditions for ozone modeling in very complex terrains: A case study in the northeastern Iberian Peninsula, *Environmental Modelling & Software*, 22 (9): 1294-1306. DOI: 10.1016/j.envsoft.2006.08.004

Jiménez P., O. Jorba, J.M. Baldasano and S. Gassó. 2008. The Use of a Modelling System as a Tool for Air Quality Management: Annual High-Resolution Simulations and Evaluation. *The Science of Total Environment* 390 (2-3): 323-340, doi:10.1016/j.scitotenv.2007.10.025.

Pay, M.T. 2011. Regional and urban evaluation of an air quality modelling system in the European and Spanish domains. Doctoral thesis. 242 pp.

Pay, M.T., Jiménez-Guerrero, P., Baldasano, J.M. 2010a. Implementation of resuspension from paved roads for the improvement of CALIOPE air quality system in Spain, *Atmospheric Environment*, 45 (3): 802-807. DOI: 10.1016/j.atmosenv.2010.10.032

Pay, M.T., M. Piot, O. Jorba, S. Gassó, M. Gonalves, S. Basart, D. Dabdub, P. Jiménez-Guerrero, J.M. Baldasano. 2010b. A full year evaluation of the CALIOPE-EU air quality modeling system over Europe for 2004. *Atmospheric Environment*, 44 (27): 3322-3342, DOI:10.1016/j.atmosenv.2010.05.040

Pay, M.T., P. Jiménez-Guerrero, O. Jorba, S. Basart, X. Querol, M. Pandolfi, J.M. Baldasano. 2012a. Spatio-temporal variability of concentrations and speciation of particulate matter across spain in the caliope modeling system. *Atmospheric Environment*, 46: 376-396, DOI:10.1016/j.atmosenv.2011.09.049

Pay, M.T., Jiménez-Guerrero, P., Baldasano, J.M. 2012b. Assessing sensitivity regimes of secondary inorganic aerosol formation in Europe with the CALIOPE-EU modeling system. *Atmospheric Environment* 51: 146-164, DOI:10.1016/j.atmosenv.2012.01.027

Pay et al. 2012c. Evaluación del sistema de pronóstico de calidad del aire CALIOPE en España.Technical Report. Available at: https://goo.gl/NbG82D.

Pay, M. T., Martínez, F., Guevara, M., and Baldasano, J. M. 2014. Air quality forecasts on a kilometer-scale grid over complex Spanish terrains, *Geosci. Model Dev.*, 7: 1979-1999, DOI: 10.5194/gmd-7-1979-2014

Schaap, M., Cuvelier, C., Hendriks, C., Bessagnet, B., Baldasano, J.M., Colette, A., Thunis, P., Karam, D., Fagerli, H., Graff, A., Kranenburg, R., Nyiri, A., Pay, M.T., Roul, L., Schulz, M., Simpson, D., Stern, R., Terrenoire, E., Wind, P. 2015. Performance of European chemistry transport models as function of horizontal resolution. *Atmospheric Environment*, 112: 90-105. DOI: 10.1016/j.atmosenv.2015.04.003

Sicardi, V., Ortiz, J., Rincón, A., Jorba, O., Pay, M.T., Gassó, S., Baldasano, J.M. 2012. Assessment of Kalman filter bias-adjustment technique to improve the simulation of ground-level ozone over Spain, *Science of The Total Environment*, 416: 329-342, DOI: 10.1016/j.scitotenv.2011.11.050

Skamarock, W., Klemp, J.B. 2008. A time-split nonhydrostatic atmospheric model for weather research and forecasting applications. *Journal of Computational Physics*, 227: 34653485.

Soret, A., JimenezGuerrero, P., Baldasano, J.M. 2011. Comprehensive air quality planning for the Barcelona Metropolitan Area through traffic management,*Atmospheric Pollution Research*, 2 (3): 255-266, DOI: 10.5094/APR.2011.032

Soret, A., JimenezGuerrero, P., Andres, D., Cardenas, F., Rueda, S., Baldasano, J.M. 2013. Estimation of future emission scenarios for analysing the impact of traffic mobility on a large Mediterranean conurbation in the Barcelona Metropolitan Area (Spain), *Atmospheric Pollution Research*, 4 (1): 22-32, DOI: 10.5094/APR.2013.003

Soret, A., Guevara, M:, Baldasano, J.M. 2014. The potential impacts of electric vehicles on air quality in the urban areas of Barcelona and Madrid (Spain), *Atmospheric Environment*, 99: 51-63. DOI: 10.1016/j.atmosenv.2014.09.048

Turco, M., Llasat, M.C., Herrera, S., Gutiérrez, J.M. 2009. Bias correction and downscaling of future RCM precipitation projections using a MOS-analog technique. *Journal of Geophysical Research*, 122 (5): 2631-2648. DOI: 10.1002/2016JD025724

Zhang, Q., Wang, H., Dong, J., Zhong, G., Sun, X. 2017. Prediction of Sea Surface Temperature Using Long Short-Term Memory. *IEEE Geoscience and Remote Sensing Letters*, 14 (10): 1745-1749. DOI: 10.1109/LGRS.2017.2733548