D4.1: Developed and tested data fusion algorithm for use in the pilot study activities WP7

WP4 – Data fusion model and reasoning services

Document Information

Grant Agreement Number	688363	Acronym	hackAIR	
Full Title	Collective awareness platform for outdoor air pollution			
Start Date	1st January 2016	Duration	36 months	
Project URL	www.hackAIR.eu			
Deliverable	D4.1 – Report on spatial mapping through data fusion			
Work Package	WP4 – Data fusion model and reasoning services			
Date of Delivery	Contractual	30 June 2017	Actual	28 June 2017
Nature	Report	Dissemination Level	Confidential	
Lead Beneficiary	NILU			
Responsible Author	Philipp Schneider (NILU)			
Contributions from	William Lahoz (NILU), Panagiota Syropoulou (Draxis)			

Document History

Version	Issue Date	Stage	Description	Contributor
1	03/05/2017	Draft	First draft	Philipp Schneider (NILU), William Lahoz (NILU)
2	21/05/2017	Draft	Added additional material or testing with real-world data	Philipp Schneider (NILU), William Lahoz (NILU)
3	03/06/2017	Draft	Fixed language and typos	Philipp Schneider (NILU), William Lahoz (NILU)
4	12/06/2017	Draft	Included additional figures	Philipp Schneider (NILU), William Lahoz (NILU)
5	16/06/2017	Draft	Included additional materia Language and other text edits	Philipp Schneider (NILU), William Lahoz (NILU)
6	21/06/2017	Final	Addressed reviewer comments	Philipp Schneider (NILU), William Lahoz (NILU), Panagiota Syropoulou (Draxis)

Disclaimer

Copyright message

Table of Contents

Executive summary

This document provides a description of the methodology we use for the data fusion module to be implemented within the framework of the hackAIR project. This module combines the observations made by volunteers in hackAIR with data from the regional version of the Copernicus Atmosphere Monitoring Service (CAMS), in order to provide a spatially exhaustive map of interpolated hackAIR observations. The outputs of the module are value-added maps of air quality for the two study sites of Germany and Norway. The maps spatially interpolate the highly uncertain information obtained from the hackAIR measurements by the volunteers using the highly accurate and complete concentration fields provided by CAMS as a spatial proxy.

In this document, we first provide the general background related to data assimilation and data fusion techniques as well as their meaning in the context of citizen science and crowdsourcing. Subsequently we present the preliminary methodology used for data fusion in hackAIR. The methodology is based on geostatistics; this allows for a mathematically meaningful interpolation of the hackAIR observations in space while at the same time inheriting the spatial patterns from the model-based concentration fields. We then briefly discuss the observational data that we expect from the hackAIR volunteers and present in more detail the modelling information that is available from CAMS at both global and regional scales and that are relevant for use in the framework of the hackAIR data fusion module. Subsequently we present mapping results using simulated observations. Results using simulated observations indicate that the method is capable of producing realistic spatial fields of air quality that on the one hand inherit the spatial patterns of the underlying model information while at the same time interpolating the gaps in the observations in a mathematically meaningful way. In addition we discuss the application of the algorithm on diverse real-world observations made by the hackAIR community and analyse the data availability in the hackAIR database at the time of writing this document. Finally, we present the general architecture and the data flow of the module and describe how it is going to be implemented on the hackAIR server.

1 Introduction

This deliverable reports on the methodology for combining the observations that will be made within the pilot studies of hackAIR with model information through data fusion, with the goal of creating real-time updated maps of air quality in the participating countries. In the following two sections, we first provide the aim of the document and describe its structure.

1.1 Aim of this document

The aim of this document is to provide a description of the data fusion module that is going to be implemented within the framework of the hackAIR project. It should be noted that the described methodology is primarily based on theoretical considerations and tests with simulated data and thus we cannot consider it as entirely complete yet at this point. Once a sufficient number of real-world observations is available on the hackAIR server, some slight changes might be made to the methodology in order to accommodate some of the characteristics of the real-world observations that could not be sufficiently evaluated with simulated observations.

1.2 Structure of the deliverable

The deliverable is structured as follows. Section 2 provides a general background on information in the Earth System. We include a description of the method of data assimilation/data fusion, and its relevance for citizen science and related concepts such as crowdsourcing as used within the framework of the hackAIR project. Subsequently, in Section 3 we provide the theoretical foundation and describe the algorithm of the suggested data fusion methodology in hackAIR. We then briefly describe the information coming from the hackAIR observations (aerosol optical depth derived from user-provided and online images as well as open hardware sensors) in Section 4, before discussing the source of operational model information we will use in hackAIR, namely the Copernicus Atmosphere Monitoring Service, in Section 5. Finally, in Section 6 we show some examples of mapping products coming out of the data fusion methodology, in this case using simulated observations. Section 7 provides an analysis of the availability of real-world observations in the hackAIR database at the time of writing this document and discusses their suitability for carrying out testing of the data fusion algorithm. Section 8 describes the technical implementation of the data fusion module on the hackAIR server, and finally Section 9 provides a short summary of the document.

2 Background

In the following, we present a broad background on information in the Earth System, data assimilation, and citizen science. Where not indicated otherwise this discussion to a large extent follows the material provided in a previous publication by the authors [*Lahoz and Schneider*, 2014].

2.1 Information

We have two broad sources of information of the Earth System: measurements, i.e., “observations”; and understanding of the spatio-temporal evolution, typically embodied in “models,” e.g., representing equations describing relationships between variables and/or parameters. Model information typically embodies our understanding of the system of interest, e.g., the Earth System. Two aspects are important regarding information about the Earth System: Firstly, both the observational and model information have uncertainty, and a key task is to understand and quantitatively estimate this uncertainty.

A second aspect of observational information is that it has spatio-temporal gaps [*Lahoz and Schneider*, 2014]. To fill in the gaps we need a model, which can be as simple as linear interpolation or complex as representing the Navier-Stokes

equations of the atmosphere. We can understand the model as an intelligent interpolator of the observational information. We would like to fill in the gaps in an objective manner, e.g., by minimizing a penalty function calculated from observational information and prior information of the system (e.g., from a model forecast). A methodology that allows this intelligent interpolation is data assimilation [*Lahoz and Schneider*, 2014]. It has strong links to several mathematical disciplines, including control theory and Bayesian estimation.

2.2 Data assimilation

The data fusion methodology applied in hackAIR is a subset of data assimilation [*Kalnay*, 2003; *Lahoz et al.*, 2010]. Data assimilation adds value to the observations by filling in the observational gaps, and adds value to the model by constraining it with observations – see Figure 1. This allows a self-consistent and realistic representation of the Earth System on a regular grid. In this way, data assimilation allows one to "make sense" of Earth Observation. In particular, data assimilation provides methods for combining in an objective way observations and models with different spatio-temporal characteristics and errors: local footprint vs. quasi-global footprint; local coverage vs. global coverage; differences in sampling frequency; and errors arising from matching different spatio-temporal scales. When we combine the observational and model information and their errors in data assimilation, we term the result the "analysis." We will never know precisely the errors in the observations, models and the analyses, so we need to estimate them. This means that we have to state the data assimilation problem in statistical terms. The weather forecasting agencies provide an example of how data assimilation combines heterogeneous observational and model information [*Kalnay*, 2003].

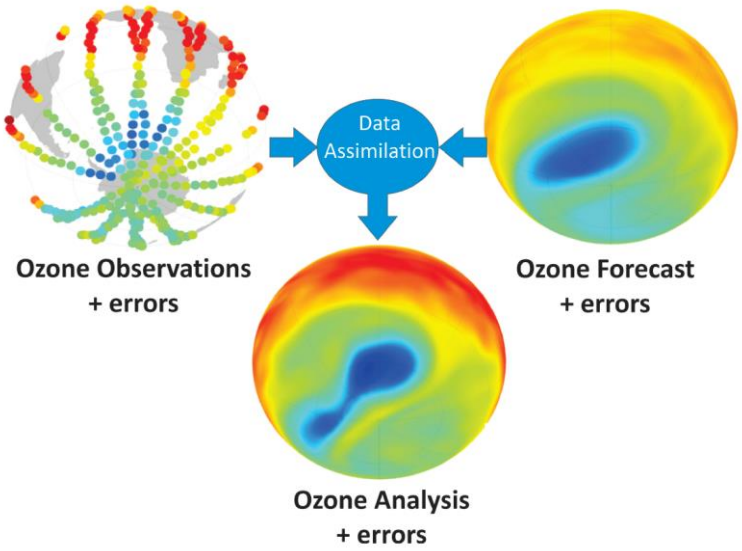

Figure 1 – Schematic of how data assimilation adds value to observational and model information. The data shown are various representations of ozone data at 10 hPa (about 30 km in altitude) on 23 September 2002. Lower left panel: plot representing ozone data from a limb-viewing satellite. Lower right panel: plot representing a 6-day forecast based on output from a data assimilation system. Top panel: plot representing an ozone analysis based on output from a data assimilation system. The analysis is produced by combination of the observational and model information and their errors. Note how the analysis fills in the observational data gaps and captures the Antarctic ozone split, verified using independent data not used in the assimilation. By contrast, the ozone hole split is not captured in the 6-day forecast. From [*Lahoz and Schneider*, 2014].

Bayesian estimation defines a systematic and rigorous approach to data assimilation [*Rodgers*, 2000]. However, its full-scale implementation in many areas, including weather forecasting, is impossible, chiefly due to the size of the problem. The typical dimension of current weather forecasting models is $\sim 10^7$ elements, while the number of observations available over 24h is $\sim 10^6$–10^7 [*Lahoz and Errera*, 2010; *Lahoz and Schneider*, 2014]. As a result, error covariance matrices for the model and observational information have $\sim 10^{14}$ elements. However, the Bayesian approach is still useful since it provides general guidelines for developing a data assimilation system and evaluating its

results. Nevertheless, in many practical applications we need to make simplifying assumptions to the data assimilation methodology. There are two main lines of work: (i) statistical linear estimation and (ii) ensemble assimilation.

In the statistical linear approach, there exist two broad classes of numerical algorithms for data assimilation: variational and sequential [*Bouttier and Courtier*, 1999]. These algorithms take respectively the form of the 4-D variational method (4D-Var), or the Kalman filter (KF). These are two different algorithms for determining the best linear unbiased estimate (BLUE) and they are equivalent only under the condition of linearity. Statistical linear estimation achieves Bayesian estimation when the system is linear and the errors are Gaussian.

The ensemble assimilation approach is a form of Monte-Carlo approximation, which attempts to estimate the error covariance matrices using a finite number of ensemble members. In the ensemble Kalman filter, EnKF [Evensen, 2003], we use a Monte-Carlo ensemble of short-range forecasts to estimate the short-term forecast error.

The major drawback of the algorithms introduced above (variational methods; sequential methods; ensemble methods such as the EnKF) is the underlying assumption that the model states have a Gaussian distribution. A modification to the KF (the extended Kalman Filter, EKF) can handle some departure from Gaussian distributions of model errors and non-linearity of the model operator (which evolves the model forward in time). However, if the model becomes too non-linear or the errors become highly skewed or non-Gaussian, the trajectories computed by the EKF will become inaccurate. A development in data assimilation using ensemble methods that addresses non-linear and non-Gaussian aspects is the particle filter, PF [*van Leeuwen*, 2009].

2.3 Citizen Science and data assimilation

Activities from citizens involved in science ("Citizen Science") provide a novel and recent development in platforms for observing the Earth System, potentially complementing the traditional ways of observing the Earth System, viz., satellite and ground-based and in situ platforms. Citizen Science activities have been described as people accumulating knowledge in order to learn about and respond to environmental threats [*Irwin*, 1995], and as public participation in scientific research [*Rosner*, 2013]. Citizen Science is closely related to similar concepts such as crowdsourcing [*Howe*, 2006; *Estelles-Arolas and Gonzalez-Ladron-de-Guevara*, 2012], participatory sensing [*Christin et al.*, 2011], and ubiquitous sensing [*De Nazelle et al.*, 2013].

While mobile air quality sensors (e.g. those used while walking or bicycling) are currently not as useful for real-time mapping purposes as static sensors due to their even higher uncertainties and the extremely high spatio-temporal variability, mobile instruments are nonetheless a very important tool in the toolbox of Citizen Science. They often rely on the fact that smartphones are increasingly ubiquitous, given growth in mobile use, changes in mobile usage, and the increasing range of features provided to mobile phone users. Through a smartphone, the citizen can provide and receive information on their immediate environment, e.g., at the most basic level using only the phone's internal sensors on temperature, noise, movement, location, or on a wide variety of other parameters using external sensor packs. This includes air quality parameters such as NO_x ($NO+NO_2$), CO, ozone, and aerosols (particulate matter, PM), measured by deploying small, low-cost external microsensors, and using a smartphone as the main communications device. In addition, smartphones allow users to provide geo-located observations on nearly any generic parameter using specific apps. There are several applications of the concept of Citizen Science. Examples include (i) the WOW (Weather Observations Website; http://wow.metoffice.gov.uk) project at the Met Office, UK – a way to obtain information on various meteorological parameters (temperature, rainfall rate, and snowfall) in the UK, and (ii) temperatures in an urban environment using solely the internal battery temperature sensors of smartphones [*Overeem et al.*, 2013]. Mobile instruments measuring air quality were tested within several projects in the past, however due to the high spatio-temporal variability of air pollution, the non-systematic fashion in which such sensors are generally handled by the public, and the difficulty of assigning an area of representativeness to such measurements, they are not ideal for the air quality mapping technique for hackAIR reported on here. We focus here primarily on open hardware sensors at static locations. It should be noted that this limitation also does not refer to

mobile AOD observations derived from pictures as it is carried out in hackAIR as these already inherently represent a spatial and temporal average and are thus less prone to small-scale variability than actual sensor measurements.

A natural path to follow with Citizen Science information is the use of data assimilation to add value to it, in the same way that it does for traditional observation platforms (see above). The use of Citizen Science for data assimilation brings its own challenges. These include the following. (i) Significantly different spatial scales compared to those at which data assimilation is traditionally performed [*Lahoz and Schneider*, 2014]. (ii) Model development, e.g., the need to simulate smaller spatial scales. (iii) Noisy information from users and from microsensors [*Shanley et al.*, 2013]. (iv) Representation of uncertainty in a way that is user-friendly and informative [*Spiegelhalter et al.*, 2011]. A further challenge is the merging of data from traditional sources such as satellite and ground-based and in situ platforms, and data provided by Citizen Science. The work carried out in hackAIR along the lines of data fusion is a first step towards overcoming some of these challenges.

3 Data fusion methodology for hackAIR

The data fusion methodology applied in the framework of the hackAIR project is based on geostatistical principles [*Isaaks and Srivastava*, 1989; *Cressie*, 1993; *Goovaerts*, 1997; *Kitanidis*, 1997; *Armstrong*, 1998; *Journel and Huijbregts*, 2003; *Wackernagel*, 2003; *Webster and Oliver*, 2007; *Sarma*, 2009; *Chilès and Delfiner*, 2012]. It uses universal kriging to combine observations with model data by predicting the concentrations at unknown location by simultaneously interpolating the observations and using the model data to provide information about the spatial patterns.

In contrast to ordinary kriging, universal kriging allows the overall mean to be non-constant throughout the domain and to be a function of one or more explanatory variables. Universal kriging is similar to kriging with external drift and mathematically equivalent to regression kriging [*Hengl et al.*, 2007] or residual kriging [*Denby et al.*, 2010; *Horálek et al.*, 2013] but can perform the linear regression against auxiliary variables and the spatial interpolation of the corresponding residuals in a single step. Universal kriging assumes a non-stationary mean and in addition the presence of local spatial variation. As such, the parameter in question is modelled by a deterministic regression component that provides the large-scale spatial variation and provides spatial patterns in areas where no observations are available, and a kriging component that provides the small-scale random variation.

In general, the estimated concentration $\hat{Y}(s_0)$ at point s_0 is computed as

$$\hat{Y}(s_0) = c + a_1 \cdot x_1(s_0) + a_2 \cdot x_2(s_0) + \cdots + a_p \cdot x_p(s_0) + \varepsilon(s_0) \quad (1)$$

Where c is a constant, a_1, a_2, etc. are regression coefficients, X_1, X_2, …, X_p are the values of the p predictor variables of the regression component, and ε is a stationary random process with a given semivariogram. In matrix notation we get

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1(s_0) & \cdots & x_p(s_0) \\ 1 & \vdots & \ddots & \vdots \\ 1 & x_1(s_n) & \cdots & x_p(s_n) \end{bmatrix} \begin{bmatrix} c \\ a_1 \\ \vdots \\ a_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix} = \boldsymbol{Xa} + \boldsymbol{\varepsilon} \quad (2)$$

where $\boldsymbol{Y}$ indicates the estimated values at all predicition locations, $\boldsymbol{X}$ represents the values of the predictor variables at all locations, $\boldsymbol{a}$ is the vector of regression coefficients, ε indicates the vector of residual errors that is estimated using kriging with the known semivariogram model, n is the number of prediction locations, and p is the number of predictor variables.

In practice, we estimate the spatial trend or drift ε of the mean using a single predictor variable, which in this case is the output of the model. In this case, there is only a single predictor variable, so Equation 1 simplifies to

$$\hat{Y}(s_0) = c + a_1 \cdot x_1(s_0) + \varepsilon(s_0). \quad (3)$$

The observations are provided by i) the hackAIR volunteers, observing aerosol optical depth (AOD) (or at least a parameter correlated with AOD) by taking dedicated photos with their phones or measuring air quality in their local neighbourhood using open hardware sensors, ii) AOD estimates from social media images (e.g. Flickr), and iii) AOD estimates from webcam images. As such, the system takes the overall spatial patterns from the model data, and dynamically adjusts this field based on the observations. The data fusion is generally carried out at the same spatial resolution at which the model data is available (in this case ca. 10 km by 10 km), however minor adjustments to the spatial scale are possible during the interpolation process. For example, in hackAIR we will produce data fusion maps with a spatial resolution of 5 km by 5 km to allow for a slightly "smoother" look of the maps.

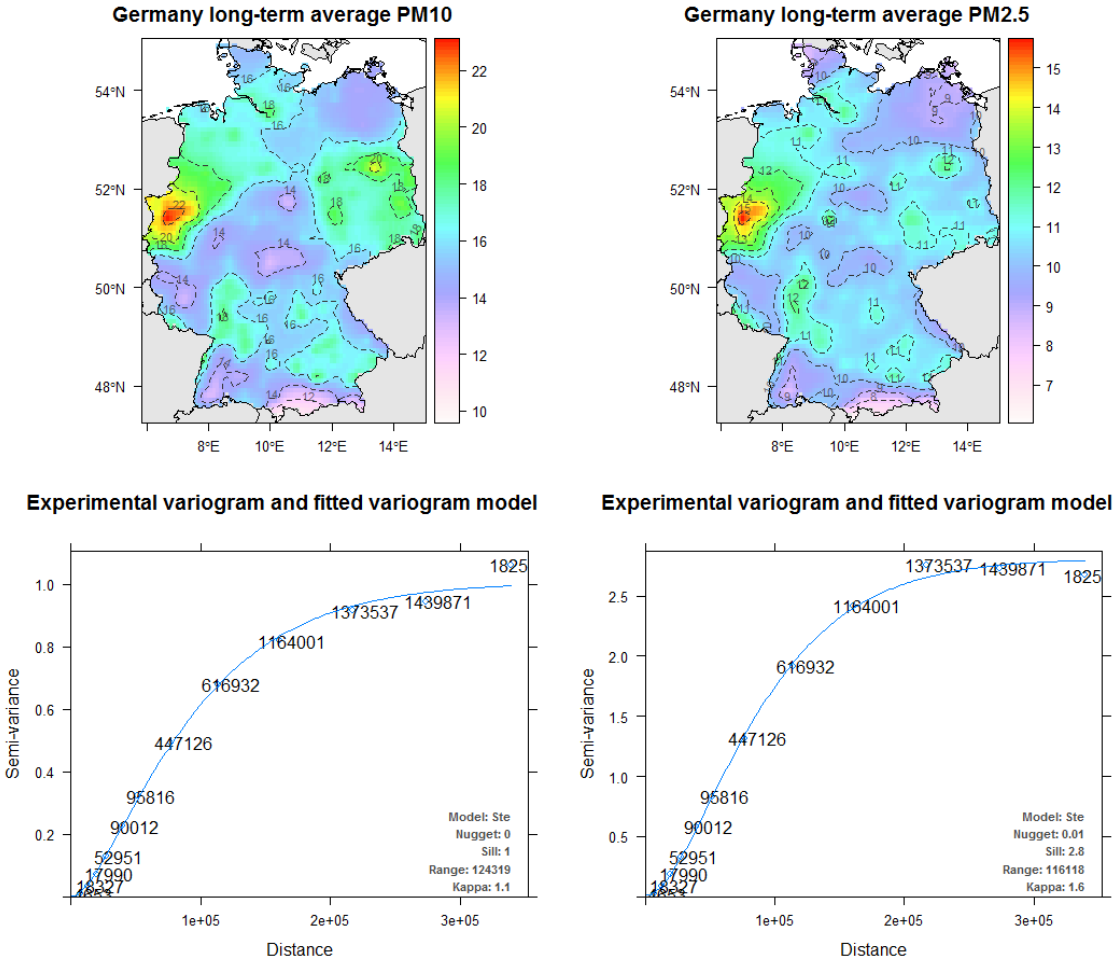

Figure 2: Model-derived long-term average maps of PM_{10}, and $PM_{2.5}$ with the corresponding semivariograms for the Germany study site.

The theoretical semivariogram required for calculating the covariances in the kriging process is fitted automatically to the empirical semivariogram for each new set of observations jointly with the respective modelled map at specified time intervals. For illustration purposes, Figure 2 shows examples of semivariograms for two species (PM_{10}, $PM_{2.5}$) derived from long-term average maps for the hackAIR study site of Germany. The semivariogram is the geostatistical method for describing the error of representativeness in that it provides information about the spatial autocorrelation structure of the observations and, if enough observations are available within the study site, can be used to draw conclusions about the area over which each individual observations is representative. As such, the hackAIR data fusion

algorithm incorporates the error of representativeness as it entirely automatically fits theoretical semivariograms to the experimental sample semivariograms derived directly from the observations.

The system can map both quantitative measurements given on a numerical scale, as well as categorical measurements observed using classes such as *low*, *medium*, and *high*. Due to their high expected uncertainties, the majority of the hackAIR measurements are expected to fall in the latter category. Nonetheless, if we use the system for numerical observations, a log transform of the data using the natural logarithm can be useful. This approach follows previous work such as that carried out by [*Denby et al.*, 2008], [*De Smet et al.*, 2010], and [*Horálek et al.*, 2014, 2015] and is done because the frequency distribution of observed and modelled concentrations most often resembles the lognormal distribution. A log-transformation, therefore, is able to convert these distributions into an approximately Gaussian distribution, which we assume for universal kriging. Taking the lognormal distribution of the concentrations into account has further been shown to provide superior mapping accuracy [*Denby et al.*, 2008; *Horálek et al.*, 2013].

If we use log-transformation of numerical observations, the resulting concentration field and the corresponding mapping uncertainty have to be back-transformed from log-space. [*Denby et al.*, 2008] showed that the theoretical back-transformed expectation value of a concentration C is given as

$$E[C] = \exp\left(\mu + \frac{\sigma^2}{2}\right) \tag{4}$$

where μ and σ represent the mean standard deviation of the log-normal transformed data, respectively. In practice, the concentration values resulting from the data fusion process are thus back-transformed by exponentiation with the kriging error as

$$\hat{Z}(s_0) = \exp\left[\hat{Y}(s_0) + \frac{\sigma^2(s_0)}{2}\right] \tag{5}$$

where $\hat{Z}(s_0)$ is the estimated back-transformed concentration value at point s_0, $\hat{Y}(s_0)$ is the concentration at point s_0 resulting from the data fusion process, and $\sigma(s_0)$ is the kriging standard deviation at point s_0 [*De Smet et al.*, 2010].

The theoretical back-transformed variance of the log-normal distribution is computed as

$$\mathrm{var}[C] = [\exp(\sigma^2) - 1] \cdot \exp[2\mu + \sigma^2] \tag{6}$$

Where μ and σ represent the mean and standard deviation of the log-normal transformed data, respectively [*Denby et al.*, 2008]. Thus, we can calculate the back-transformed standard deviation (uncertainty) $\delta(s_0)$ at point s_0 of the fused map in practice as

$$\delta(s_0) = \sqrt{\exp[(\sigma^2(s_0) - 1] \cdot \exp[2 \cdot \hat{Y}(s_0) + \sigma^2(s_0)]} \tag{7}$$

Where $\sigma(s_0)$ is the kriging standard deviation at point s_0 and $\hat{Y}(s_0)$ represents the concentration at point s_0 resulting from the data fusion process [*Denby et al.*, 2008; *De Smet et al.*, 2010].

4 Observational data from hackAIR

The hackAIR project will provide observations from two primary data sources operated by the volunteers recruited within the two hackAIR pilot studies.

The first set of observations comes from the algorithm that analyses the red-to-green ratio of the sky in photographs in order to derive a quantity resembling aerosol optical depth. The photographs to be used will primarily come from online websites such as Flickr (https://www.flickr.com/), which provide a large number of images that are uploaded

very quickly after they have been taken. Similarly to image hosting sites such as Flickr, pictures from sky-depicting webcams will be exploited using the same methodology. While suitable webcams are relatively rare and thus are not expected to offer much in terms of spatial sampling, they have the tremendous advantage that they provide a continuous stream of data with sampling rates ranging from seconds to at least hourly. They also are advantageous over random online images as they are mounted in a fixed location and always provide pictures of the same scene using the same camera with a constant calibration. This reduces a lot of the sources of uncertainty compared to images acquired from other sources. In addition to online images and webcams, recruited hackAIR volunteers will provide sky-depicting photos through the hackAIR app. We expect that the quality of these sky-depicting images will be significantly better than that of the acquired online images; however, their number likely will be significantly lower on a daily basis.

The second set of observations in hackAIR will come from open hardware sensors that will be distributed to volunteers throughout Germany and Norway (the two hackAIR study sites). We expect that for operational purposes the AOD product retrieved from sky-depicting images will be more valuable for the data fusion as a) their number (and thus their spatial density) will be significantly larger than that of open hardware sensors and b) they will be available in a continuous stream of data and not limited to occasional measurement campaigns by volunteers. For this reason, the data fusion module has been primarily developed for fusing the information from sky-depicting images at the country-scale. However, the information from open hardware sensors will also be included in the data fusion process although technically their information is more representative for the very local intra-urban scale. In addition to the observations collected within the framework of the hackAIR project, there are other sources of information about air quality. Most notably this includes observations by air quality monitoring stations utilizing reference equipment. This data is included here indirectly as the data fusion module uses up to date daily forecasts from the CAMS regional models, which assimilate station observations of air quality and thus ensure that the resulting concentration fields are corrected by observations.

5 Model information: CAMS

In order to spatially interpolate the observation from the hackAIR community, information from a spatially exhaustive model is required in order to guide the interpolation. In this section, we discuss the Copernicus programme and the Copernicus Atmospheric Monitoring Service (CAMS), which serves as the source of modelling data for the hackAIR data fusion module. CAMS is currently the most mature operational modelling system for atmospheric composition worldwide and provides crucial modelling information used in the data fusion module together with the actual observations from user-provided and online images as well as open hardware sensors collected within the framework of the hackAIR project.

5.1 Copernicus

Copernicus is a programme designed for the establishment of a European system for monitoring the Earth and is coordinated and managed by the European Commission. The goal of the programme is to provide a long-term operational set of systems to collect Earth Observation data through satellites and in situ measurements (ground-based stations, airborne and seaborne platforms), to process the resulting data and to provide the resulting up-to-date information to end users via a number of services related to environmental and security issues. Copernicus is organized within two main components, the Copernicus Space Component and the Copernicus Services. Figure 3 and Figure 4 illustrate the respective organizational structure of the Copernicus Space Component and Copernicus Services Component.

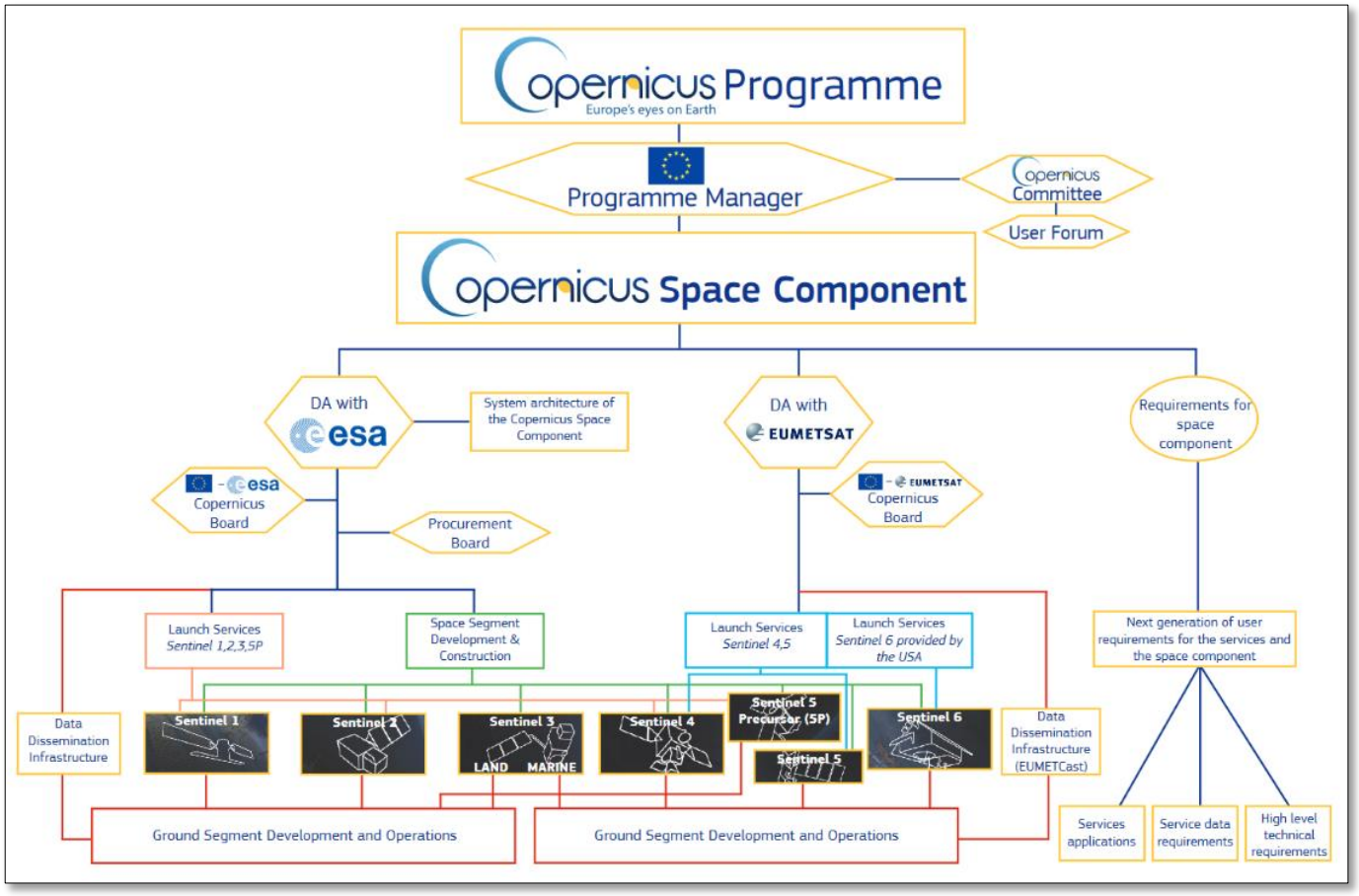

Figure 3 - Overview of the organizational structure of the Copernicus Space Component (Figure taken from the website www.copernicus.eu)

The Space Component of Copernicus, which is coordinated by the European Space Agency (ESA), includes dedicated earth observation satellites (Sentinel-1, -2, -3, -5P, and -6) and instrumentation onboard of EUMETSAT's weather satellites (Sentinel-4, and -5). In addition, the Space Component includes contributing missions that are operated by national, European, and international organizations.

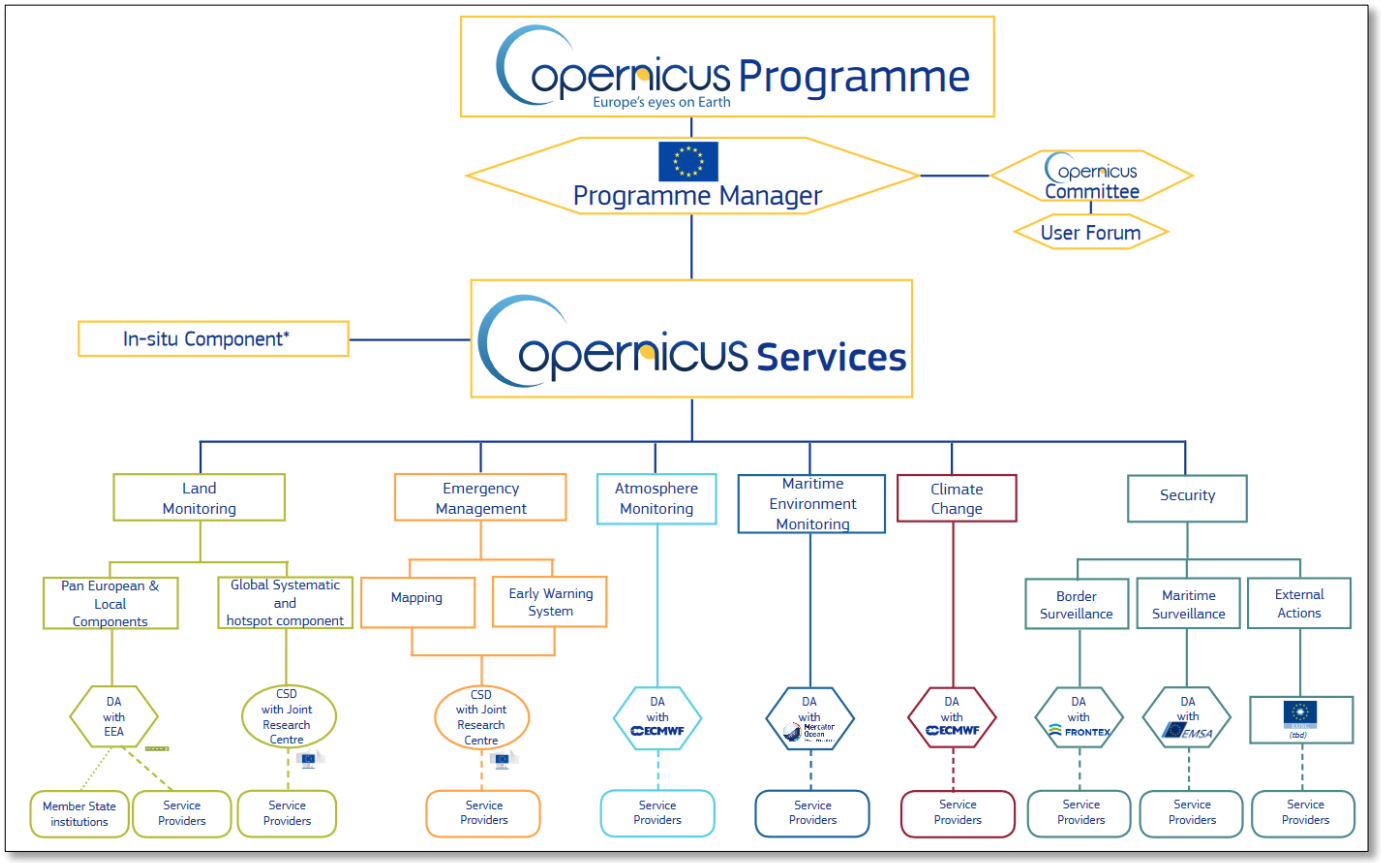

Figure 4 - Overview of the organizational structure of the Copernicus Services Component (Figure taken from the website www.copernicus.eu)

The Copernicus Services address six thematic areas: atmosphere, climate change, land applications, marine applications, emergency management and security. The primary end users of Copernicus data are considered to be policymakers and public authorities. The Copernicus services can give such stakeholders important information for developing environmental legislation and allow for critical decision-making in times of crises and emergencies.

The Service of the most relevance for the hackAIR project is the Copernicus Atmosphere Monitoring Service (CAMS). This products provided by this service are available under the website `http://atmosphere.copernicus.eu/`. CAMS is described in detail in the following Section.

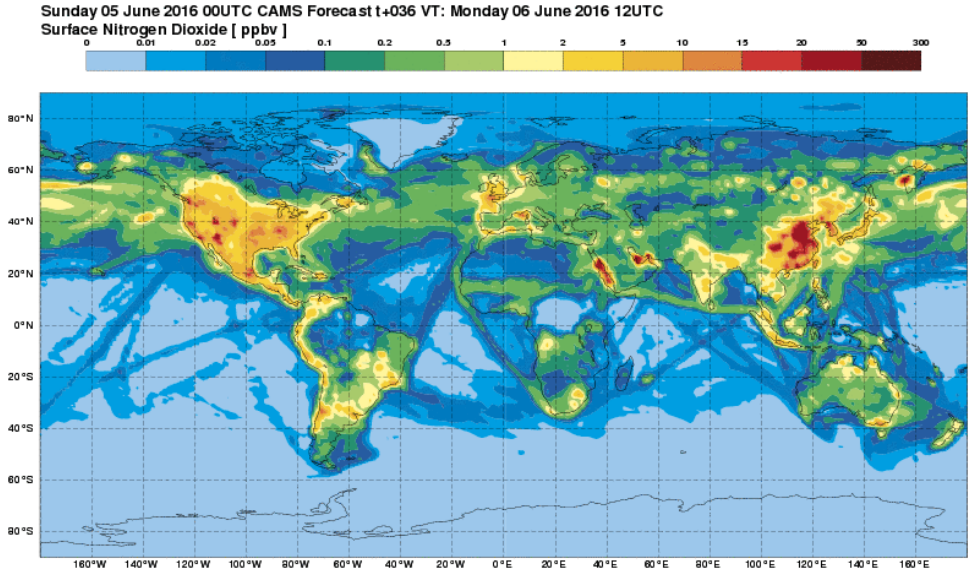

Figure 5 - Example of global output from CAMS as produced by the Integrated Forecast System (IFS), here showing the forecast for surface-level nitrogen dioxide for Monday June 6 2016 at 12 UTC.

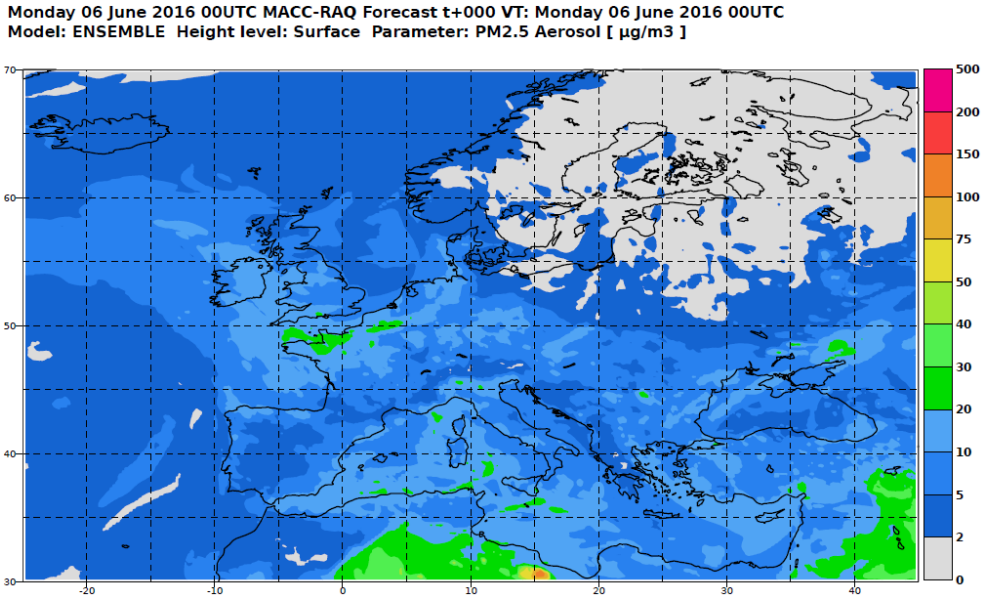

Figure 6 - Example of regional output from CAMS as produced as the ensemble of seven different European-scale models, here showing the forecast for surface-level PM2.5 for Monday 6 June 2016 at 0 UTC.

5.2 Copernicus Atmosphere Monitoring Service (CAMS)

CAMS consolidates many years of preparatory research (carried within a series of projects called Monitoring of Atmospheric Composition and Climate, MACC, MACC-II, MACC-III) and development and delivers the following operational services (this list is provided by ECMWF, 2016):

- Daily production of near-real-time analyses and forecasts of global atmospheric composition;
- Reanalyses providing consistent multi-annual global datasets of atmospheric composition with a frozen model/assimilation system;
- Daily production of near-real-time European air quality analyses and forecasts with a multi-model ensemble system;
- Reanalyses providing consistent annual datasets of European air quality with a frozen model/assimilation system, supporting in particular policy applications;
- Products to support policy users, adding value to "raw" data products in order to deliver information products in a form adapted to policy applications and policy-relevant work;
- Solar and UV radiation products supporting the planning, monitoring, and efficiency improvements of solar energy production and providing quantitative information on UV irradiance for downstream applications related to health and ecosystems;
- Greenhouse gas surface flux inversions for CO_2, CH_4 and N_2O, allowing the monitoring of the evolution in time of these fluxes;
- Climate forcings from aerosols and long-lived (CO_2, CH_4) and shorter-lived (stratospheric and tropospheric ozone) agents.

The products delivered by CAMS can be grouped into three main service lines, comprising the global, regional (Europe) and value-added products, which are directly derived from the former or relate to atmospheric composition variables (these lists are provided by ECMWF, 2016).

Main global products:

- Near-real-time analyses and forecasts for aerosol, reactive and greenhouse gases, stratospheric ozone and related species, and UV radiation;
- Delayed-mode analyses of aerosol and greenhouse gases;
- Reanalyses of aerosol, reactive gases, greenhouse gases and stratospheric ozone and consistent aerosol direct and indirect forcing.

Main regional products (Europe: 25°W-45°E, 30°N-70°N):

- Near-real-time analyses and forecasts of air quality from an ensemble of regional systems;
- Seasonal pollen forecasts;
- Annual regional air quality reanalyses.

Main value-added products:

- Global and European anthropogenic emissions inventories;
- Daily global emissions from wildfires;
- Delayed-mode surface fluxes of carbon dioxide, methane, and nitrous dioxide;
- Reanalyses of global carbon dioxide and methane surface fluxes;
- Surface solar irradiance;
- Policy-oriented products (assessment reports; country and region source-receptor calculations; "green scenario" forecasts);
- Products in support of special events and scientific campaigns.

Figure 5 and Figure 6 shows examples for the global and regional output from CAMS, respectively.

5.3 Regional CAMS datasets

For the country-scale applications as they are envisioned within the framework of the hackAIR project, it is primarily the regional (European) products offered by CAMS that are of interest. The CAMS regional system is a set of seven state-of-the-art air quality models that were developed in Europe and focus, therefore, on the European domain (30°N - 70°N, 25°W - 45°E) at relatively high spatial resolution (varying from model to model between ~10 km and ~20 km). The models used within the CAMS regional system include

- CHIMERE developed by INERIS, France [*Schmidt et al.*, 2001];
- EMEP developed by MET Norway [*Simpson et al.*, 2003];
- EURAD-IM developed by the University of Cologne, Germany [*Hass et al.*, 1995];
- LOTOS-EUROS developed by KNMI and TNO, Netherlands [*Schaap et al.*, 2008];
- MATCH developed by the SMHI, Sweden [*Andersson et al.*, 2015];
- MOCAGE developed by Meteo-France, France [*Josse et al.*, 2004];
- SILAM developed by FMI, Finland [*Sofiev et al.*, 2008].

All these regional models are run using identical input data with regard to meteorology (obtained from the ECMWF global weather forecasting system), boundary conditions for chemical species (acquired from the CAMS IFS-MOZART global production), and emissions (obtained from the CAMS emission dataset for anthropogenic emissions over Europe as well as biomass burning).

Each individual model every day produces 4-day forecasts with hourly temporal sampling. In addition, all models perform a daily retrospective analysis of the pollutants near the surface. These products assimilate ground-based station data from the last 24 hours to constrain the model results.

List of forecast products currently available (png/pdf for plotted data and GRIB2/NetCDF for numerical data):

Species	O3	NO2	SO2	CO	PM10	PM2.5	NH3	NO	NMVOC	PANs	Birch pollen
Forecast Level: surface	●	●	●	●	●	●	●	●	●	●	●
Forecast Levels: from surface up to 5000m	●	●	●	●	●	●	●	●	●	●	●
plotted data (PNG/PDF) Level: surface	●	●	●	●	●	●	●	●	●	●	●
Online data (NetCDF)	●	●	●	●	●	●	●	●	●	●	●
Archived data (GRIB2)	●	●	●	●	●	●	●	●	●	●	●

Figure 7 - List of individual model forecast products currently available from the CAMS regional production system (from: http://regional.atmosphere.copernicus.edu/)

List of analysis products currently available (png/pdf for plotted data and GRIB2/NetCDF for numerical data):

Species	O3	NO2	SO2	CO	PM10	PM2.5	NH3	NO	NMVOC	PANs	Birch pollen
Analysis Level: surface	●	●	●	●	●	●	●	●	●	●	●
plotted data (PNG/PDF) Level: surface	●	●	●	●	●	●	●	●	●	●	●
Online data (NetCDF)	●	●	●	●	●	●	●	●	●	●	●
Archived data (GRIB2)	●	●	●	●	●	●	●	●	●	●	●

Figure 8 - List of individual model analysis products currently available from the CAMS regional production system (from: http://regional.atmosphere.copernicus.edu/)

List of Ensemble products currently available (png/pdf for plotted data and GRIB2/NetCDF for numerical data):

Species	O3	NO2	SO2	CO	PM10	PM2.5	NH3	NO	NMVOC	PANs	Birch pollen
Forecast Level: surface	●	●	●	●	●	●	●	●	●	●	●
Forecast Levels: from surface up to 5000m	●	●	●	●	●	●	●	●	●	●	●
Analysis Level: surface	●	●	●	●	●	●	●	●	●	●	●
plotted data (PNG/PDF) Level: surface	●	●	●	●	●	●	●	●	●	●	●
plotted data (PNG/PDF) Levels: 500m, 1000m or 3000m	●	●	●	●	●	●	●	●	●	●	●
Online data (Grib2/NetCDF)	●	●	●	●	●	●	●	●	●	●	●
Archived data (GRIB2)	●	●	●	●	●	●	●	●	●	●	●

Figure 9 - List of ensemble products currently available through the CAMS regional production system (from: http://regional.atmosphere.copernicus.edu/)

In addition to the individual model runs, the CAMS regional system also uses the output from the seven regional models to generate ensemble products from both the forecasts and analysis products [*Marécal et al.*, 2015]. Furthermore, the near-real time operational runs are complemented by validated multi-model ensemble reanalysis products that are further offered for historical periods.

The main access point for the CAMS regional products as of summer 2016 is the website http://www.regional.atmosphere.copernicus.eu/. Figure 10 shows a screenshot of the web interface of the CAMS regional data portal. This data portal provides information about the offered services and can be used to administrate operational FTP-push tasks for data products. The products are available either through individual requests or through a subscription service offering FTP-push. Figure 7 through Figure 9 provide an overview of which individual species are available at each product level.

It should be noted that CAMS regional modelling system does not currently provide aerosol optical depth (AOD) as an operational output, so no direct fusion of the hackAIR-provided AOD estimates will be possible at this point. This limitation of the CAMS regional service could be compensated with the global CAMS model output (IFS-MOZART), which provides an operational AOD product and is as of summer 2016 being run at a spatial resolution of 40 km (improved from previously 80 km), however this is still very coarse for the applications envisioned in hackAIR. The solution for this lack of high-resolution AOD model information is to use the operational high-resolution $PM_{2.5}$ concentration fields (at 1 degree by 1 degree spatial resolution, or ca. 10 km by 10 km) to act as a proxy and guide the spatial interpolation of the hackAIR AOD measurements. AOD is closely related to $PM_{2.5}$ and the spatial patterns between the two parameters will be quite similar at the country-level scales considered here. The data fusion technique used here primarily uses the spatial patterns of the model information. In addition the $PM_{2.5}$concentration fields will be very suitable for data fusion of the PM measurements to be made by open hardware sensors by the hackAIR volunteers.

Figure 11 shows an example of the hourly forecast of the CAMS regional ensemble product of $PM_{2.5}$ for a 24-hour period, in this case for the region of southern Norway for 12 July 2016. This is the dataset we will use to provide the modelling information for the hackAIR data fusion module. Figure 12 shows another example of the CAMS regional ensemble product of $PM_{2.5}$, in this case for the area of Germany for October 27 2016.

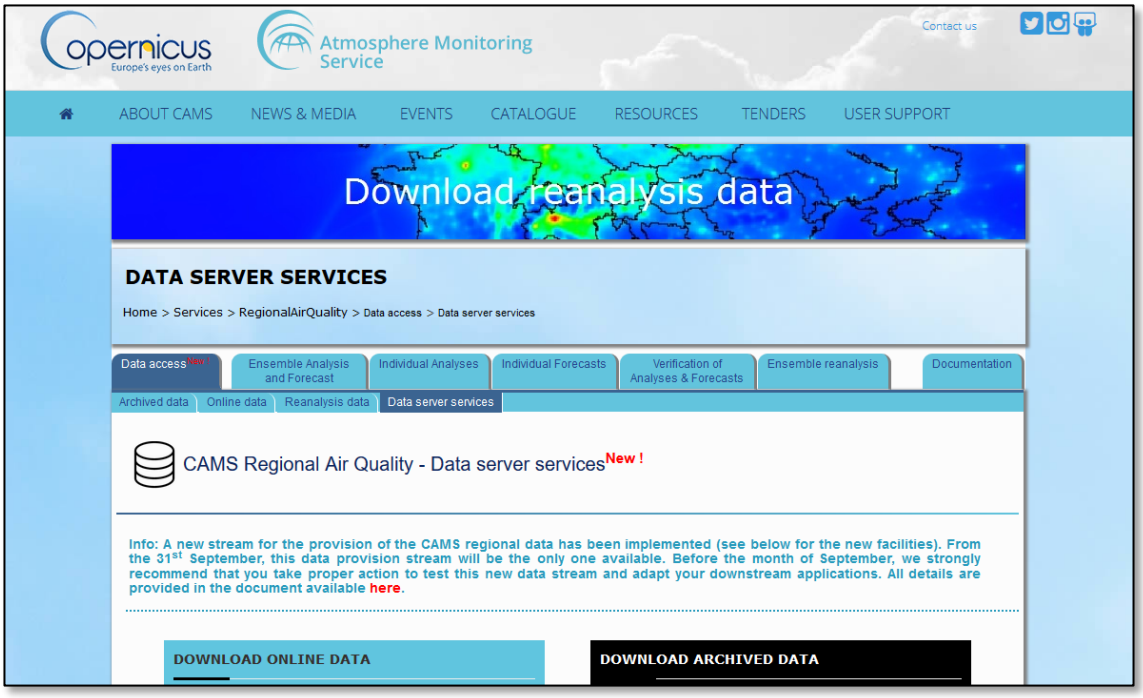

Figure 10 - Web interface of the CAMS regional data portal.

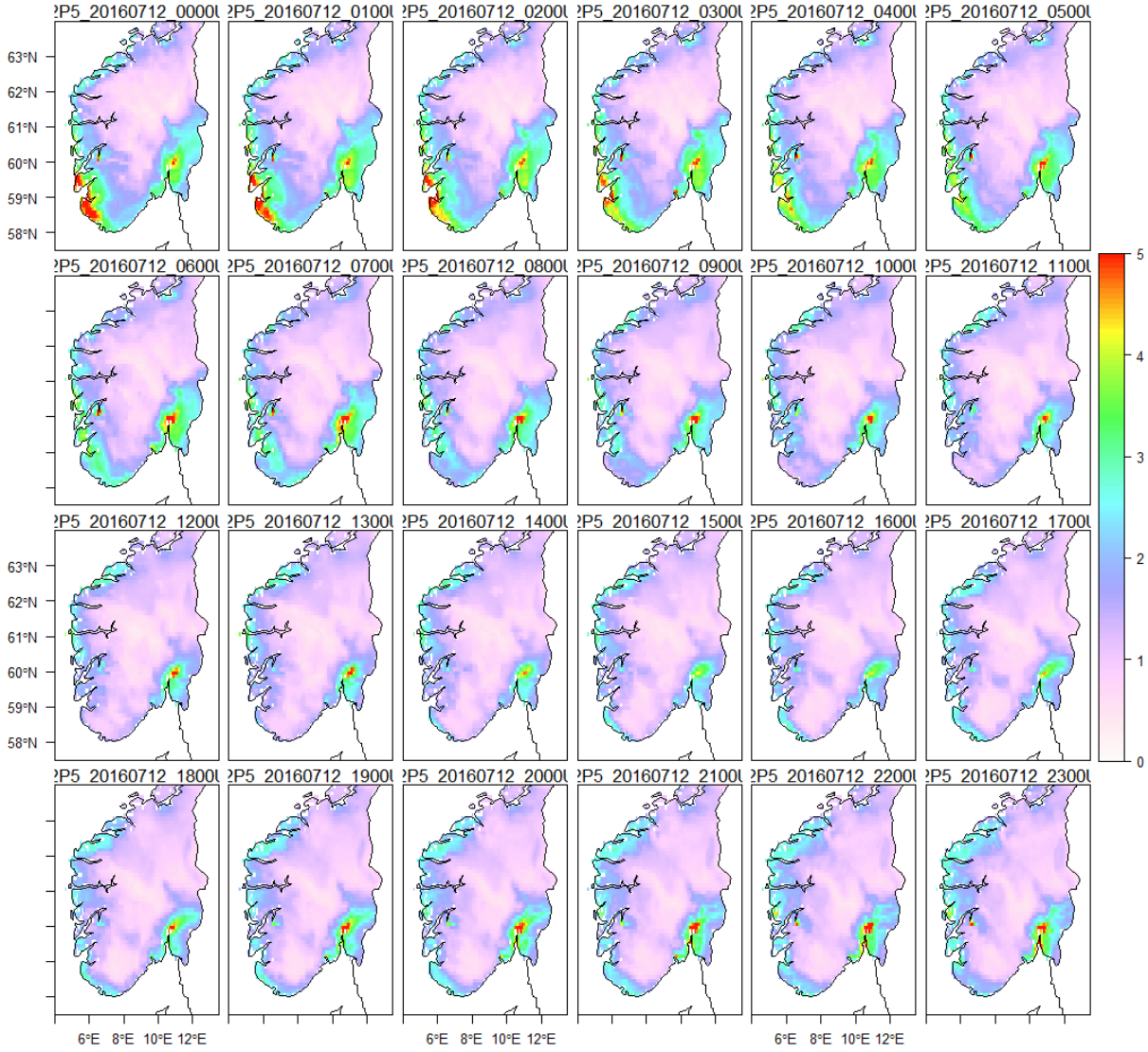

Figure 11 - Hourly CAMS regional ensemble forecast data of PM2.5 (in units of µg/m^3) for a period of 24 hours, here shown for the area of southern Norway on July 12 2016.

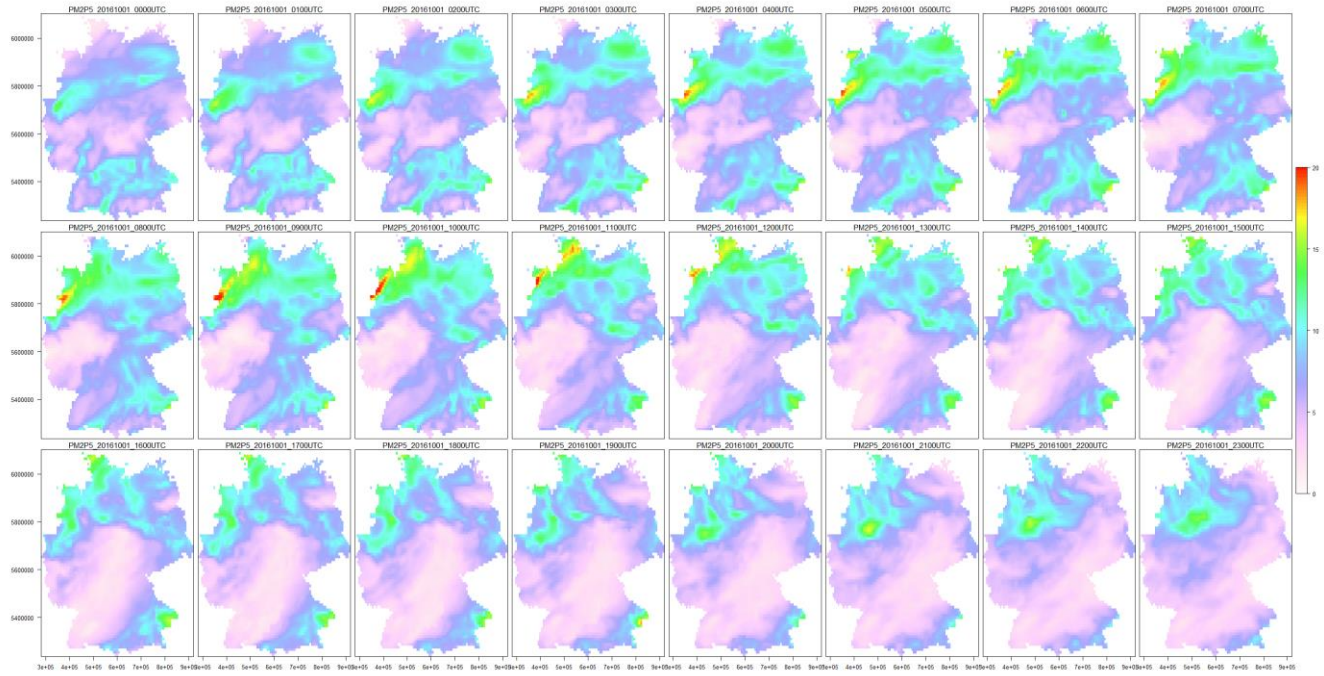

Figure 12: Hourly CAMS regional ensemble forecast data of PM2.5 (in units of μg/m^3) for a period of 24 hours, here shown for the area of Germany on October 27 2016.

5.4 Other model information

It was decided early on in the project to focus primarily on data fusion at the country level and that operational CAMS modelling information would be the ideal dataset for this purpose. However, given a large enough network of accurate air quality sensors, the hackAIR data fusion methodology can in principle also be carried out at the urban scale. In fact this has already been demonstrated in other projects such as the FP7 project CITI-SENSE [*Schneider et al.*, 2017]. For urban-scale data fusion, only very accurate sensors can be considered as they need to be able to represent the street-level concentrations. If, after thorough evaluation, the hackAIR open hardware sensors are found to be able to reproduce official PM values in real-world conditions and if the Norway pilot study of hackAIR will manage to deploy a network of ca. 50 operational open hardware sensors within the city of Oslo, we are planning to demonstrate the data fusion methodology in the Oslo area with hackAIR data.

For this purpose we will make use of the EPISODE model [*Slørdal et al.*, 2003], which is a combined 3D Eulerian/Lagrangian air pollution dispersion model for urban and local-to-regional scale applications. The model is typically used to calculate air pollution concentration in cities and urban areas from several simultaneous emission sources such as road traffic, domestic (home) heating and industry. The model calculates ground level hourly average concentrations as gridded values (using one or more user defined grids) and/or at individually placed receptor points. The model also calculates hourly dry and wet deposition values for the same geographical locations. Since the output from the model consists of hourly data, it can be used as a basis for calculating long term concentration averages or total deposition values. It also contains a statistical module for calculating the N highest daily or hourly values during the simulation period which can be used for defining percentiles. The Eulerian part of the EPISODE model consists of the numerical solution of the atmospheric (mass) conservation equation of the pollutant species in a three-dimensional Eulerian grid. The Lagrangian part of the model consists of separate subgrid-models for line- and point-sources. The line source model is an integrated Gaussian type of model, while the point source model is a Gaussian segmented plume/puff trajectory model. The meteorological data which are being used in EPISODE is calculated in a separate meteorological preprocessor. This meteorological preprocessor is based on advanced atmospheric boundary layer similarity theory. Calculations of NO_2 are based on using photochemical equilibrium between the three fast-cycle

compounds NO, NO_2 and O_3. For more comprehensive photochemical calculations, the model contains a newly developed and simplified photochemistry scheme for cities and urban areas. This scheme is based on the more comprehensive EMEP photochemistry scheme.

As the EPISODE model is not run in an operational fashion like for example CAMS, the urban-scale data fusion carried out in Oslo uses an annual average concentration map representing the typical spatial patterns of air pollution. Figure 13 shows the 2013 annual average concentration map of PM_{10} for the city of Oslo as it was derived using the EPISODE dispersion model. The basemap for Oslo derived from the EPISODE model has a spatial resolution of 100 m. This is thus also the spatial resolution at which the city-scale data fusion would be carried out for Oslo.

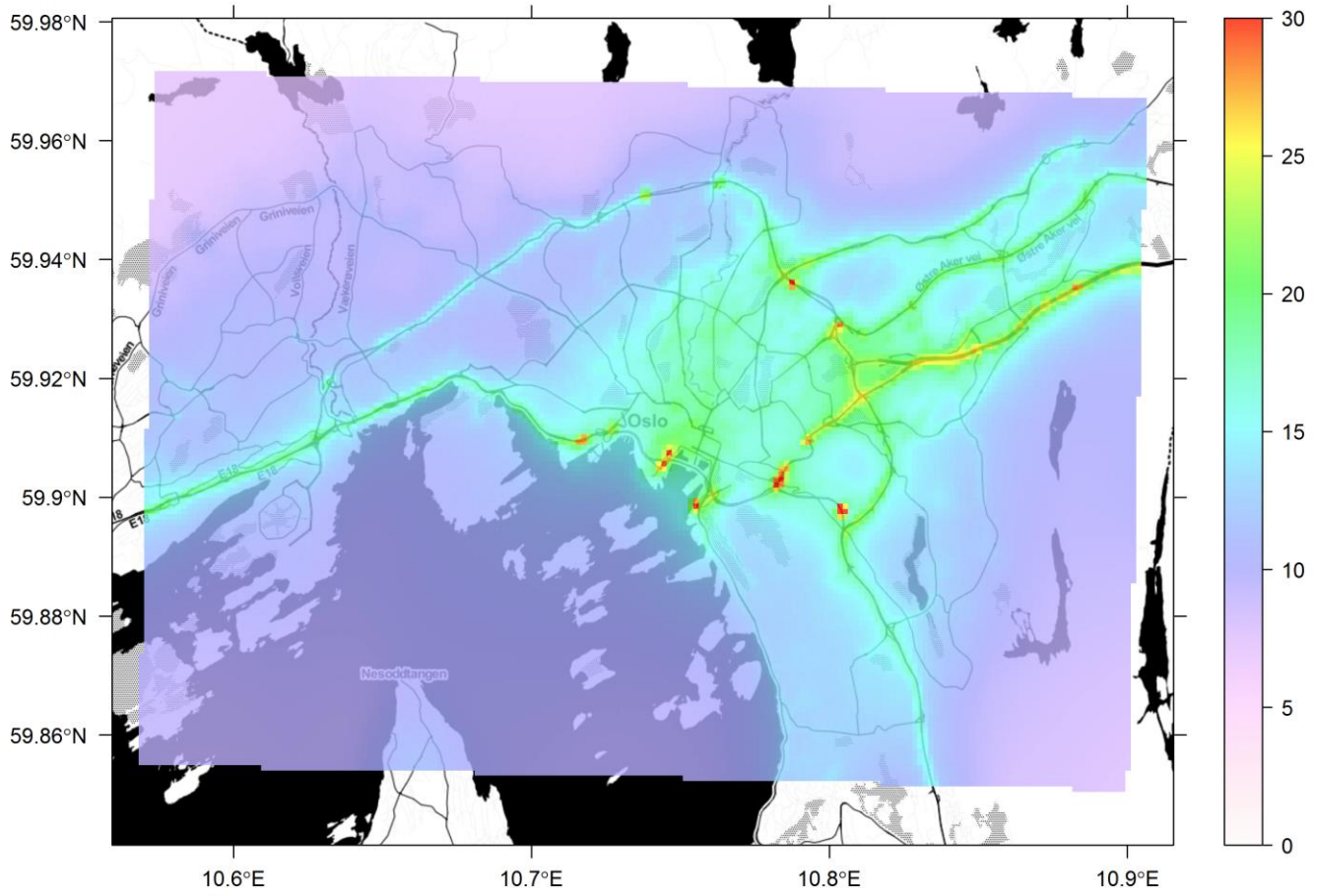

Figure 13 - Example of annual average concentration map (a "basemap") for PM_{10} for the Oslo area derived using the EPISODE dispersion model.

6 Results using simulated data

As no real observations from hackAIR volunteers were available for developing and testing the data fusion algorithm at the beginning of the hackAIR project, as realistic as possible observations were simulated in order to develop, implement, and test the data fusion algorithm. Simulating artificial observations from a model-derived concentration field, which is assumed to represent the ”true” state of the atmosphere, has the advantage that the applied mapping methodology can be tested thoroughly and validated against the same model-derived ”true” concentration field. It should be noted that of course the true state of the atmosphere is never known in practice and that this technique is

only used for algorithm development as it allows to directly study the impact of changes in the algorithm on its performance.

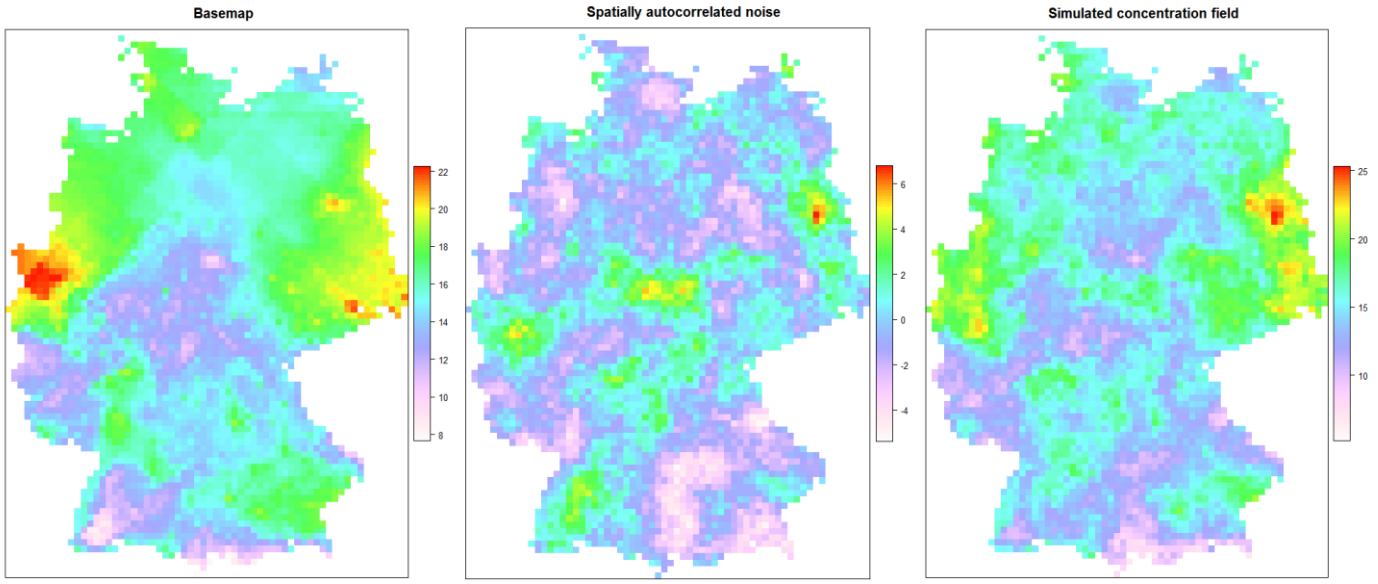

Figure 14 - First part of the workflow for simulating observations as they would be expected from the hackAIR volunteers, here shown using the example of the Germany study site. The left panel shows a long-term average maps of concentration (PM10 in this case for illustrative purposes). Independent of the basemap, a spatially autocorrelated noise field with a given mean and variance is generated using unconditional Gaussian sequential simulation (center panel). Finally, the long-term average concentration field and the spatially autocorrelated noise field are added in order to result in the final simulated concentration field (right panel).

The simulated observations were generated in a multi-step process (see Figure 14 and Figure 15 for an example illustrating the principle for the Germany study site). First, modelling information was used to generate a long-term average concentration field (a ”basemap”). Second, a field containing spatially autocorrelated noise was generated using unconditional sequential Gaussian simulation [*Goovaerts*, 1997] using a specified variance and semivariogram model. This noise field is then added to the long-term average concentration field to generate a simulated concentration field. Subsequently, a set of points distributed randomly in space is overlaid over the simulated concentration field and the simulated concentration value is sampled at each point. Finally, the concentration values at each point are converted into categorical classes as they are expected from the observations made by the hackAIR community. This system is realistic enough for testing purposes without being overly complex. One drawback of the simulation procedure at this point is that the observations locations are distributed more or less evenly in space. In reality, some clustering of the observations in the vicinity of large urban areas would be expected. A future version of the simulation algorithm might take this effect into account, however for initial testing the current method provides quite reasonable results.

Once a suitable model dataset and a realistic set of potential hackAIR observations is available, the data fusion can be carried out. Figure 16 shows an example of the process. The universal kriging is applied to the numerical equivalent of the categorical observations (i.e., Class ’low’ is defined as 1, Class ’medium’ is defined as 2, etc.). The dataset from the model, which acts as the predictor variable in the universal kriging framework, is taken at its original numerical values. Based on this procedure we can interpolate the observations classes using the modelled concentration field as a proxy for the underlying spatial patterns.

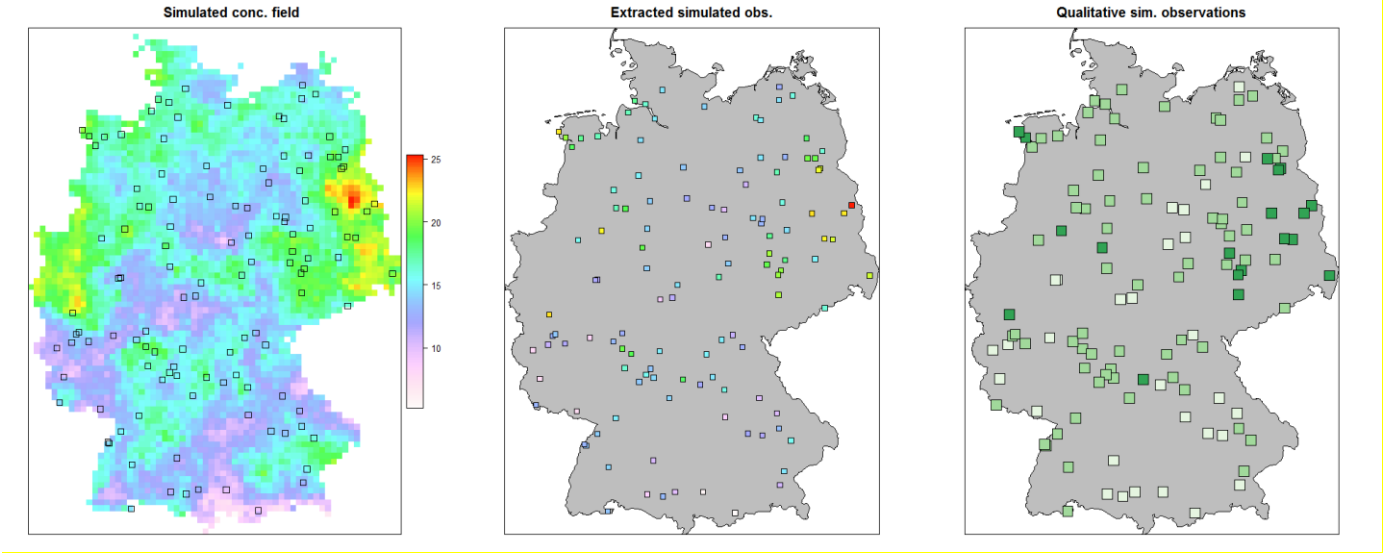

Figure 15 - Second part of the workflow for simulating observations as they would be expected from the hackAIR volunteers, here shown using the example of the Germany study site. The simulated concentration field (see also Figure 14) is shown in the background of the left panel. A random set of observation locations is the distributed over the concentration field (square markers in the left panel) and we extract the concentration at each observation location from the simulated concentration field using bilinear interpolation. The result of this process is shown in the centre panel where the location as well as the simulated concentration at each observation site are given. Finally, as an optional step, one can convert the simulated observed concentrations into a categorical class system, which is more representative of the kind of observations that are expected to be made by the volunteers within the framework of the hackAIR project (right panel).

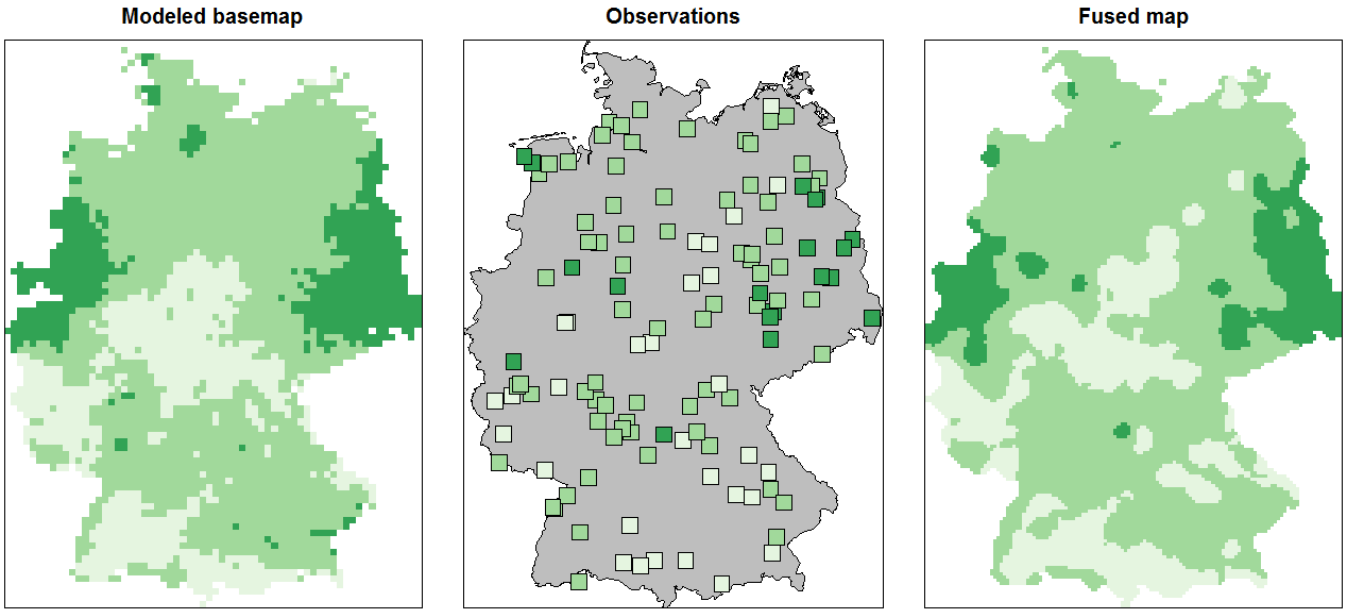

Figure 16 - Example of the categorical data fusion using simulated data, here shown for the hackAIR study site of Germany. The left panel shows a modelled concentration field (essentially similar to the CAMS-derived fields we will use in the real-world application). The centre panel shows the simulated hackAIR observations simulated using the methodology outlined in Figure 14 and Figure 15. The right panel shows the resulting fused map that combines the modelled field with the simulated observations and thus adds value to both input datasets by correcting the model information with real measurements while at the same time interpolating the actual observations in a mathematically meaningful way. Datasets like the one shown in the right panel will form the basis for a web mapping service to be implemented on the hackAIR server.

Once the data fusion is carried out, we can compare the result to the simulated "true" concentration field. This is what Figure 17 shows in an example. Here we can see how well the combination of model and observations was capable of reproducing an "unknown" truth field. The right hand panel of the figure shows the difference in class between the fused map and the truth field (after conversion from actual concentration to classes). We see that the algorithm performed well for approximately 70% to 80% of the area of the study site (when the difference is equal to zero). Areas that show a class difference are generally due to an entire lack of observation or at least to an insufficient number of them. In order to see this, compare the right hand panel of Figure 17 with the centre panel of Figure 16.

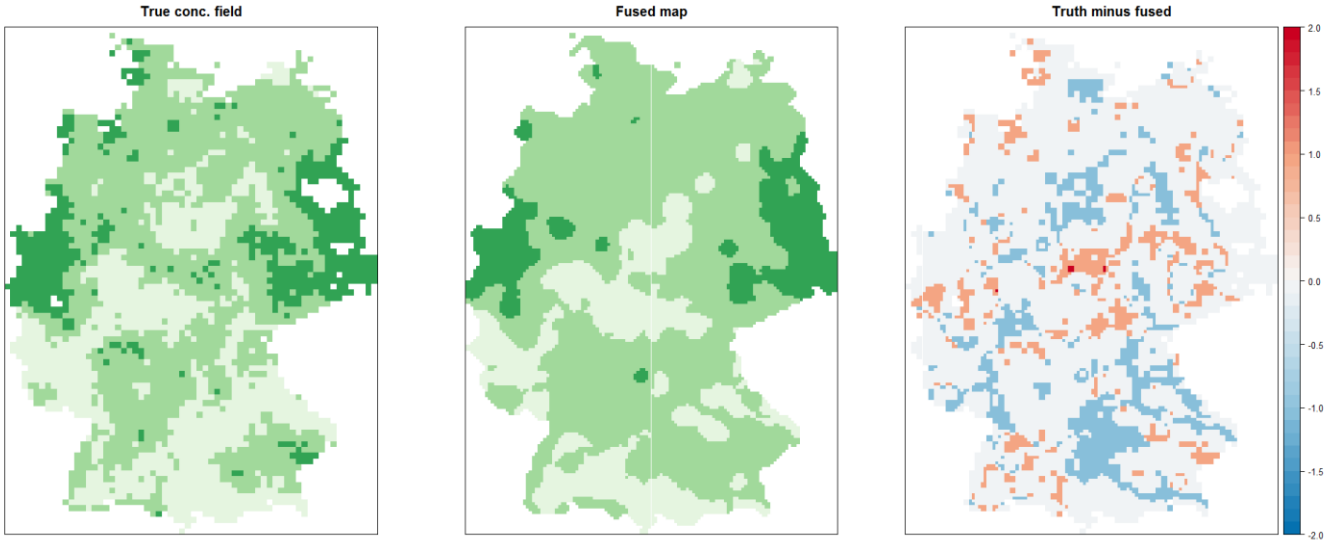

Figure 17 - Comparison of the result of the data fusion with the simulated "true" concentration field. The left panel shows a "true" concentration field simulated using the methodology presented above. The centre panel shows the results of the data fusion with simulated observations. The right panels shows the differences in classes between the data fusion result. We can observe that the algorithm does a good job in predicting the right class in about 70% of the area of the study sites (difference is equal to zero). In areas where there is a class difference, this mainly relates to the fact that no observations were available in that region.

7 Results using real-world hackAIR observations

Geostatistical algorithms, such as the universal kriging algorithm applied here generally require the calculation of an empirical semivariogram and the fit of a theoretical semivariogram model to this data. In order to do this, a minimum number of observations for each study site is necessary for properly characterizing the semivariogram. The minimum number required for this is dependent on several factors and the recommended minimum values vary somewhat throughout the geostatistical literature but a general consensus is approximately 50 valid observations distributed throughout the entire study site [*Isaaks and Srivastava*, 1989; *Goovaerts*, 2008; *Chilès and Delfiner*, 2012]. In addition, other experiments with data fusion carried out previously [*Schneider et al.*, 2017] have indicated that the mapping uncertainty becomes acceptable at a number of ca. 50 observations. It should be also noted that the number of observations required for deriving a realistic semivariogram increases with the uncertainty of the observations. If the observations themselves are highly uncertain, a larger number of observations is necessary to calculate a realistic semivariogram than if the observations are quite certain.

At the time of writing this document, the hackAIR database is beginning to be slowly populated with actual real-world hackAIR observation stemming primarily from aerosol estimates derived from online Flickr images. Some observations

from initial tests with the open hardware sensors is also available in the database but at this point only in a very limited amount.

As the population of the database with hackAIR observations has just started recently, the total number of observations for each study site is currently significantly below the number required for applying geostatistical data fusion techniques. Figure 18 and Figure 19 shows the historical availability of hackAIR observations for each over the last few months for the Germany and Norway study sites, respectively. For the Germany study site we can observe that, when there is data available, the total number of observations throughout the entire day (each 24 hour period) lies usually somewhere between 1 and 5. Only in mid-May was there a short period during which the number of observations exceed 5 for several days and even reached a maximum of 25 observations on May 21st. This is a promising result given that many more data sources will be ingested into the hackAIR database over the next few months and thus the number of usable data points is likely to increase significantly.

For the Norway study site, the situation looks somewhat more problematic. Due to the significantly lower population living there as compared to the Germany study site, the number of Flickr images uploaded there are quite low and thus for the majority of days no observations are available and at maximum one to two observations per day (each 24 hour period) are available. However, CERTH has recently shown that Norway actually has one of the highest number of available webcam sites throughout all of Europe, and thus such a data source could contribute significantly to increase the number of usable observations in Norway to levels that are sufficient for interpolating them with the data fusion algorithm.

Figure 20 and Figure 21 show the spatial distribution of the available observations for a subset of the entire period (here for mid-May as this is period where the most observations are available for the Germany study site). Looking specifically at Figure 20 we can observe that the spatial distribution of the observations occurs relatively at random and not with too much clustering. This is promising as such a spatial distribution allows for much easier processing by the data fusion algorithm. However, it can also be seen once again that the overall number of observations is still quite low as compared to what was used for testing the algorithm with simulated observations.

This situation is likely to change quite soon when the hackAIR database will be populated with data from a) a larger number of Flickr images (i.e. also some of those images that are not officially geotagged by the users but there location can still be inferred from the metadata of the image), b) images from webcams, c) user-provided images using the hackAIR app c) open hardware sensors.

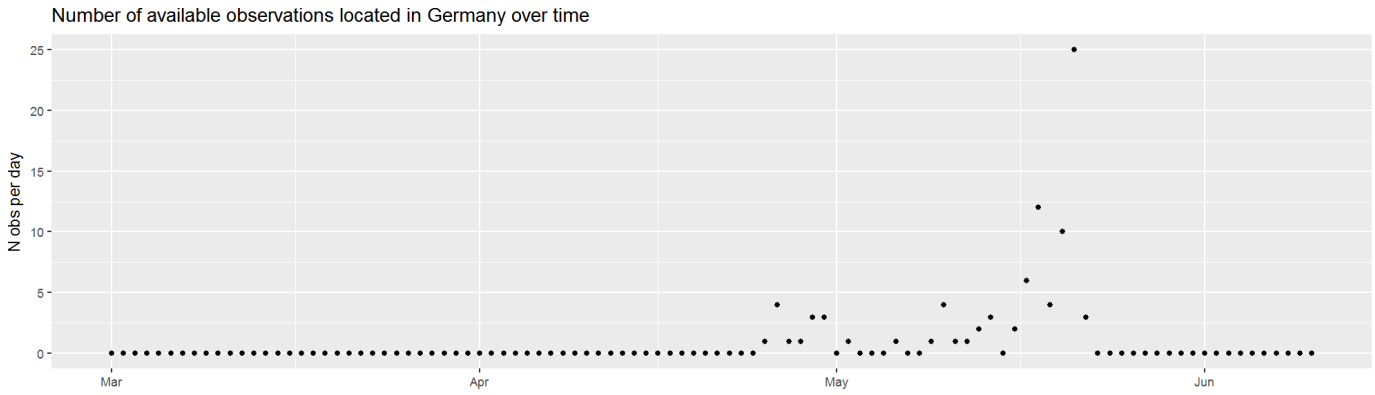

Figure 18: Time series of the number of observations located in the Germany study site available in the hackAIR database for each day.

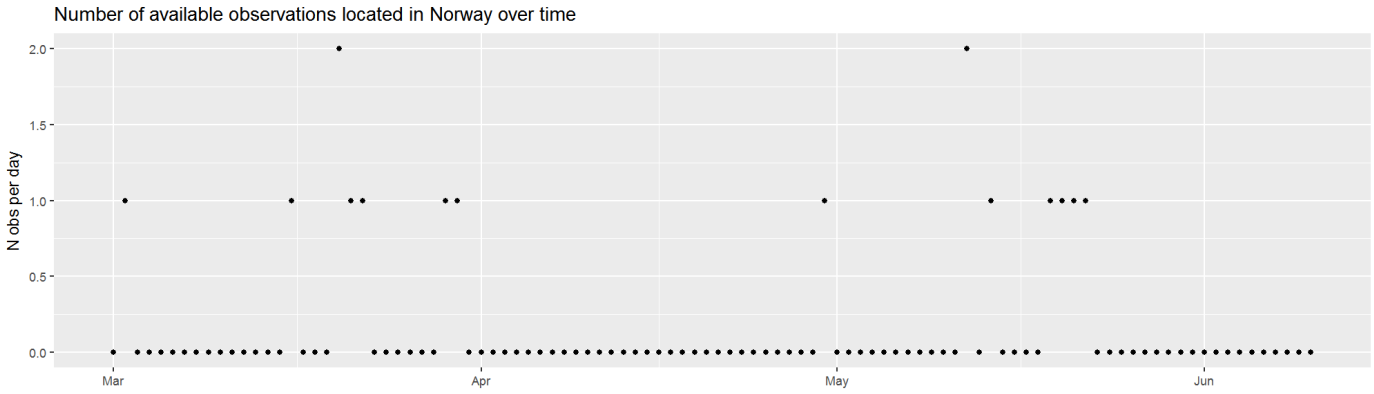

Figure 19: Time series of the number of observations located in the Norway study site available in the hackAIR database for each day.

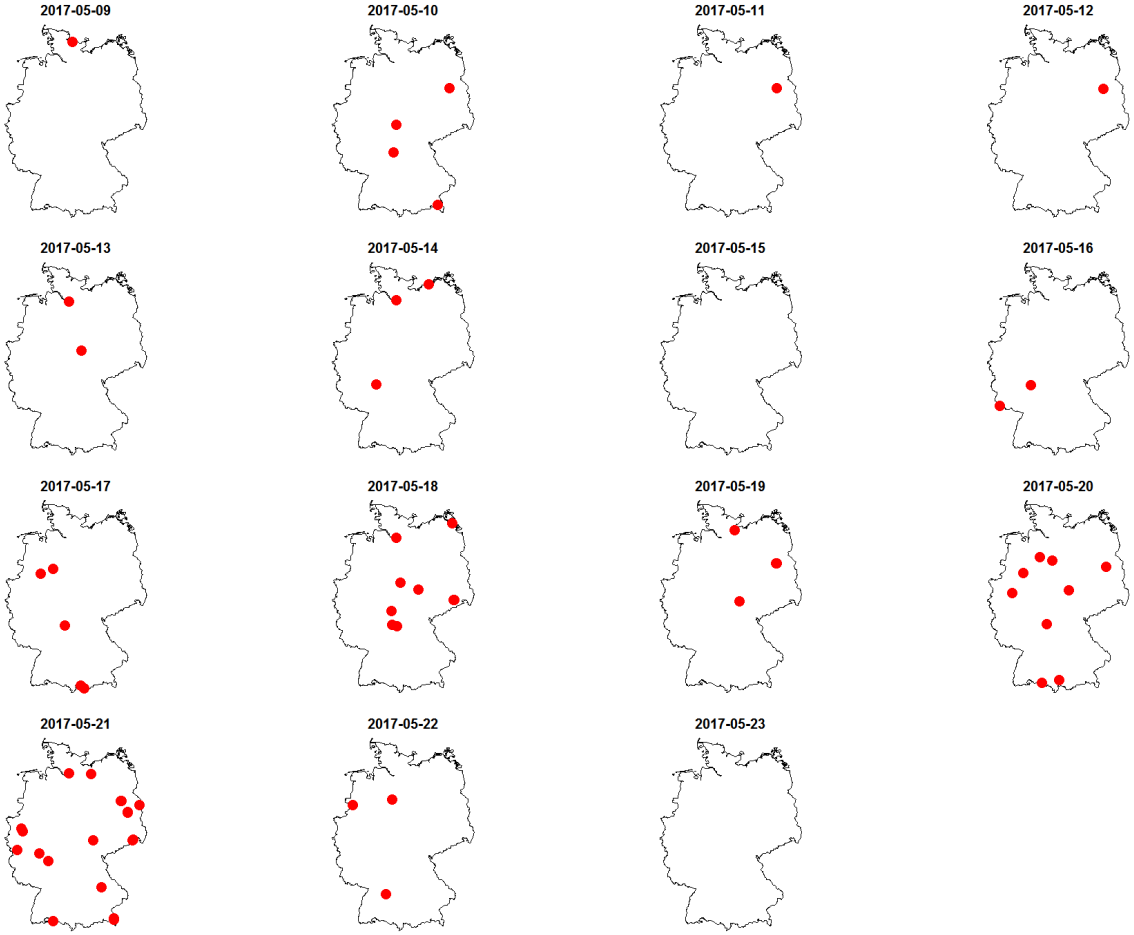

Figure 20: Daily maps illustrating the currently typical number and spatial distribution of hackAIR observations in the database, here shown for the Germany study site for a period in mid-May.

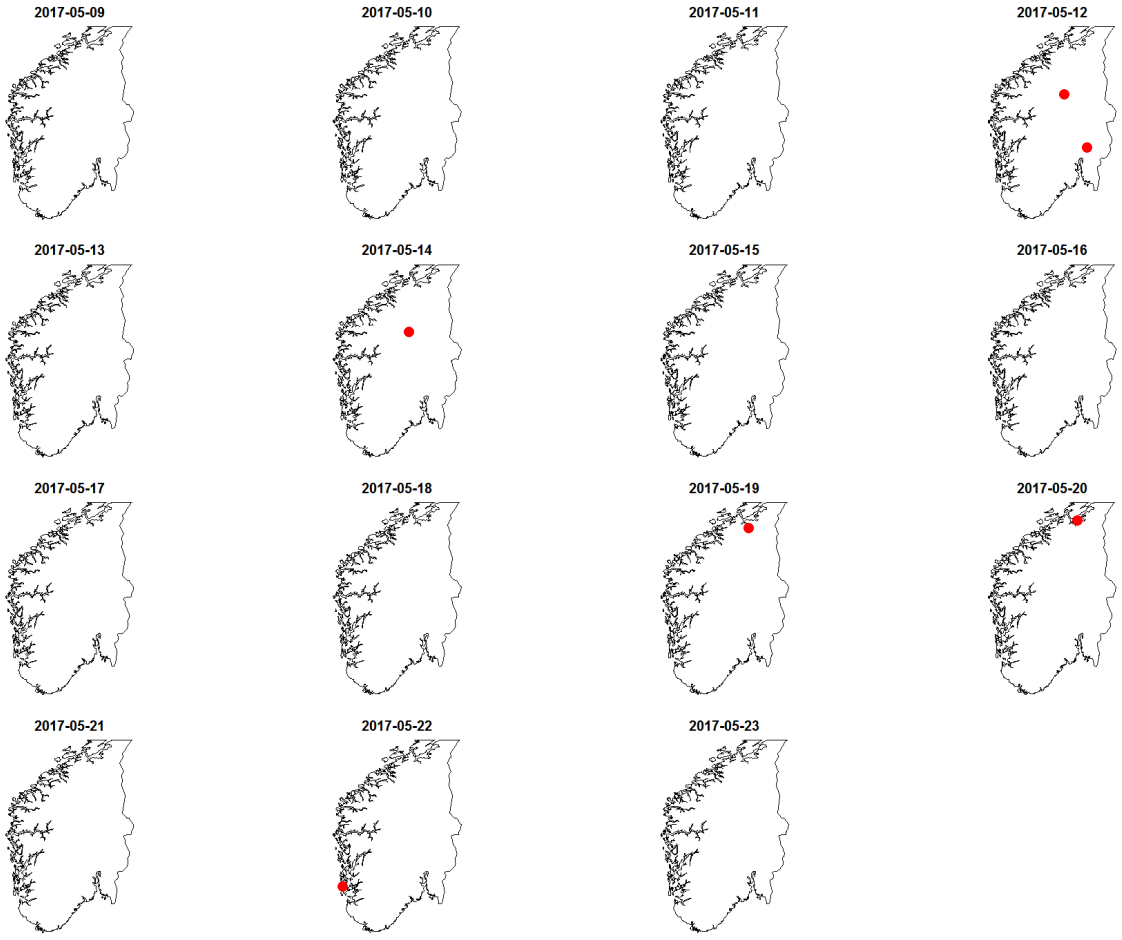

Figure 21: Daily maps illustrating the currently typical number and spatial distribution of hackAIR observations in the database, here shown for the Norway study site for a period in mid-May.

8 Technical implementation

The technical implementation of the data fusion component of hackAIR is as follows. The main data fusion code is written in the R programming language [*R Core Team*, 2016]. The R version used at the time of writing this document is R version 3.3.2 (2016-10-31). In addition, a variety of R packages are used, most notably the `sp` package (version 1.2-4) for providing spatial foundation classes [*Bivand et al.*, 2013], the `rgeos` package [*Bivand and Rundel*, 2016] version 0.3-23 for geospatial operations, the `gstat` package version 1.1-5 [*Pebesma*, 2004] for performing the universal kriging and the `automap` package version 1.0-14 [*Hiemstra et al.*, 2009] for automated fitting of the semivariogram. The main R function for the data fusion model will be regularly called on the hackAIR server using a cronjob or similar Unix tool. The timing depends on the averaging period for the data which in turn depends on the number of measurements that will be available for a given study site during the averaging period. As a general rule of thumb, the data fusion will be able to operate with a minimum of approximately 50 observations throughout the study site. If ideally at least 50 or more comparable hackAIR observations (i.e., at least 50 observations from user-generated and online images or at least 50 observations from the open hardware sensor) are regularly available for each hour,

the averaging period can be set to the minimum of one hour. This is the highest frequency at which the CAMS regional ensemble forecasts are generally available. If the number of comparable and useable hackAIR observations coming into the database is significantly lower than this for each hour, we can simply extend the averaging period and thus the mapping frequency to multiple-hour intervals or to daily averages. Based on the number of observation ingested in the database currently, the most likely outcome is that the data fusion will be run once per day.

Figure 22 shows a general overview of the architecture and data flow of the data fusion component of hackAIR. The data fusion module receives data from two main sources. Firstly, on the left we see the Copernicus Atmospheric Monitoring Service (CAMS), from which the data fusion module receives the data in NetCDF format via a REST API. The data fusion module retrieves the daily CAMS multi-model ensemble forecast after it becomes available at approximately 06:30 UTC. This product contains hourly forecast fields for all hackAIR study sites of all relevant species (here we use primarily the concentration fields for particulate matter). On the right side of Figure 22 we see the hackAIR database. In this database, the hackAIR observations coming from the user-provided and online images and from the open hardware sensors are stored. Each time the data fusion module runs, it requests the relevant observations for the desired period (with the period length depending on the number of available hackAIR observations). The observations in this period are then combined with the CAMS data as outlined in Section 2 and the resulting maps for both study sites (Germany and Norway) are then stored on the hackAIR server in the form of geoTIFF files. While the data fusion generally is carried out at the same spatial resolution as the model data (in our case this is ca. 10 km by 10 km), these output datasets will be produced at ca. 5 km by 5 km to provide slightly more spatial detail. These files can then be used by the hackAIR web portal for displaying these data using a suitable web mapping services and thus to disseminate the information to the users.

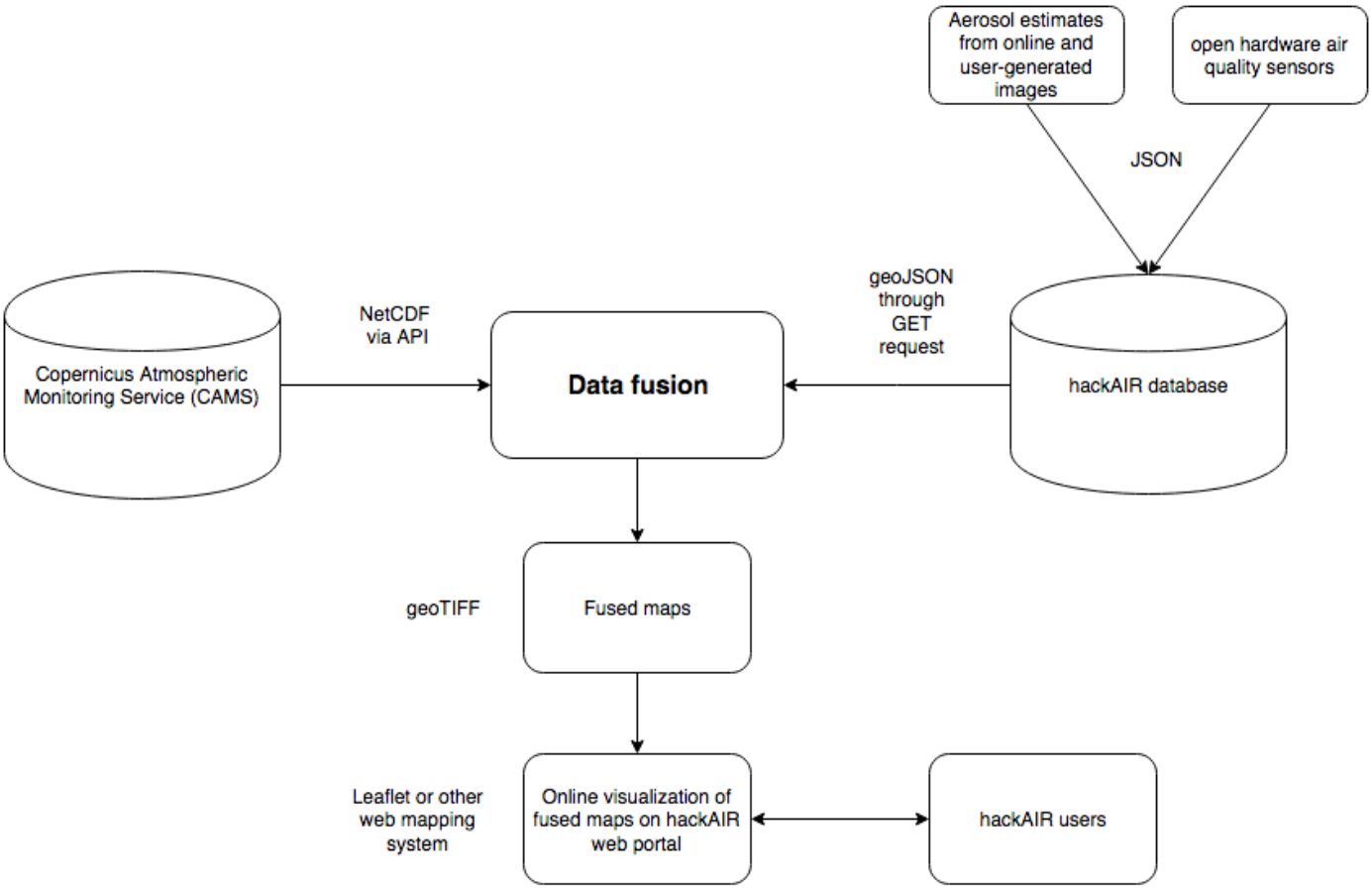

Figure 22 - General architecture and data flow of the hackAIR data fusion module

9 Conclusions

This document provides a brief overview of the data fusion and mapping methodology that is in the process of being implemented for the hackAIR project.

After providing a general background on data assimilation techniques and citizen science, we briefly describe the general methodology we use for spatially merging the information obtained from CAMS with the information provided by the hackAIR observations coming from user-provided and online images as well as open hardware sensors. The method is based on geostatistical techniques and is able to provide the best linear unbiased estimator (BLUE). More specifically, we use universal kriging to interpolate the observations made by hackAIR users using the up-to-date hourly CAMS forecasts as a spatial proxy (predictor variable). The system can operate with both quantitative measurements given on a numerical scale as well as categorical measurements observed using classes such as *low*, *medium*, and *high*. Due to the high uncertainties associated with them, we expect the majority of the hackAIR measurements to fall in the latter category. Nonetheless, if the system can be used for numerical observations, in which a log transform of the data using the natural logarithm can be useful.

We also introduce the Copernicus Atmosphere Monitoring Service here in general and provide information about its specific products at the global and regional level. CAMS provides important operationally available information for the high-resolution mapping tasks to be carried out within the project. In particular, the CAMS regional ensemble forecast at a spatial resolution of 0.1 degree by 0.1 degree (or ca 10 km by 10 km) is a crucial dataset to provide up-to-date spatial model maps for the data fusion process.

We then provide some examples of the methodology applied to simulated observations and historical modelling data. Using simulated data is important for algorithm development as it provides the opportunity for comparing the resulting concentration fields against a known reference (or “truth” field). Having knowledge of such a reference dataset allows for calculating accurate metrics quantifying the success of the methodology and is thus able to contribute significantly to testing of the resulting product as well as to further improve the underlying algorithms.

Subsequently we provide a short analysis of the feasibility of testing the data fusion algorithm with real-world observations collected by the hackAIR community. The results indicate that at the time of writing this document the hackAIR database was not populated with enough observations yet (as the pilot studies have not started) to carry out testing of the data fusion methodology with real-world data. However the outlook in terms of data acquisition is promising and over the next few months significantly larger amounts of observations from various data source will be ingested into the database and will thus enable more in-depth testing of the data fusion methodology with real-world observations before the data fusion module will be used in the pilot studies.

Finally, this document provides an overview of the technical implementation of the data fusion module we will implement within the framework of the hackAIR project. The module relies on a set of inter-linked functions written in the R program language and further relies on several open-source R packages related to spatial data handling, geostatistics, and raster processing.

Overall, the preliminary data fusion methodology presented here is able to provide automated maps that spatially interpolate the hackAIR observations made by the hackAIR community with the help of highly accurate modelling results from the operational Copernicus Atmosphere Monitoring Service. As such, the resulting datasets are able to provide the hackAIR volunteers and the interested public with a frequently updated interactive map that shows them in a user-friendly and non-scientific fashion what the air quality looks like throughout their entire country in general and in their broad region in particular. In addition, the maps that will be shown to the users are not just a model-derived scientific

map but are directly influenced by the set of observations that the users make themselves within the framework of the project. The maps therefore allow for the volunteer's direct involvement and thus contribute to awareness raising of air quality issues within the public.

10 References

Andersson, C., R. Bergström, C. Bennet, L. Robertson, M. Thomas, H. Korhonen, K. E. J. Lehtinen, and H. Kokkola (2015), MATCH-SALSA - Multi-scale Atmospheric Transport and CHemistry model coupled to the SALSA aerosol microphysics model - Part 1: Model description and evaluation, *Geosci. Model Dev.*, *8*(2), 171–189, doi:10.5194/gmd-8-171-2015.

Armstrong, M. (1998), *Basic linear geostatistics*, Springer.

Bivand, R. S., and C. Rundel (2016), rgeos: Interface to Geometry Engine - Open Source (GEOS),

Bivand, R. S., E. Pebesma, and V. Gómez-Rubio (2013), *Applied Spatial Data Analysis with R*, Second edi., Springer.

Bouttier, F., and P. Courtier (1999), *Data assimilation concepts and methods March 1999*.

Chilès, J.-P., and P. Delfiner (2012), *Geostatistics: Modeling Spatial Uncertainty*, John Wiley & Sons.

Christin, D., A. Reinhardt, S. S. Kanhere, and M. Hollick (2011), A survey on privacy in mobile participatory sensing applications, *J. Syst. Softw.*, *84*(11), 1928–1946, doi:10.1016/j.jss.2011.06.073.

Cressie, N. A. C. (1993), *Statistics for spatial data*, Wiley-Interscience, New York.

Denby, B., M. Schaap, A. Segers, P. Builtjes, and J. Horálek (2008), Comparison of two data assimilation methods for assessing PM10 exceedances on the European scale, *Atmos. Environ.*, *42*(30), 7122–7134, doi:10.1016/j.atmosenv.2008.05.058.

Denby, B., I. Sundvor, M. Cassiani, P. de Smet, F. de Leeuw, and J. Horálek (2010), Spatial mapping of ozone and SO2 trends in Europe., *Sci. Total Environ.*, *408*(20), 4795–806, doi:10.1016/j.scitotenv.2010.06.021.

ECMWF (2016), Copernicus Atmosphere Monitoring Service, Available from: http://www.ecmwf.int/en/about/what-we-do/copernicus/copernicus-atmosphere-monitoring-service (Accessed 6 June 2016)

Estelles-Arolas, E., and F. Gonzalez-Ladron-de-Guevara (2012), Towards an integrated crowdsourcing definition, *J. Inf. Sci.*, *38*(2), 189–200, doi:10.1177/0165551512437638.

Evensen, G. (2003), The Ensemble Kalman Filter: theoretical formulation and practical implementation, *Ocean Dyn.*, *53*(4), 343–367, doi:10.1007/s10236-003-0036-9.

Goovaerts, P. (1997), *Geostatistics for natural resources evaluation*, Oxford University Press, New York.

Goovaerts, P. (2008), Kriging and semivariogram deconvolution in the presence of irregular geographical units, *Math. Geosci.*, *40*(1), 101–128, doi:10.1007/s11004-007-9129-1.

Hass, H., H. J. Jakobs, and M. Memmesheimer (1995), Analysis of a regional model (EURAD) near surface gas concentration predictions using observations from networks, *Meteorol. Atmos. Phys.*, *57*(1–4), 173–200, doi:10.1007/BF01044160.

Hengl, T., G. B. M. Heuvelink, and D. G. Rossiter (2007), About regression-kriging: From equations to case studies, *Comput. Geosci.*, *33*(10), 1301–1315, doi:10.1016/j.cageo.2007.05.001.

Hiemstra, P. H., E. J. Pebesma, C. J. W. Twenhöfel, and G. B. M. Heuvelink (2009), Real-time automatic interpolation of ambient gamma dose rates from the Dutch radioactivity monitoring network, *Comput. Geosci.*, *35*(8), 1711–1721, doi:10.1016/j.cageo.2008.10.011.

Horálek, J., P. De Smet, P. Kurfürst, F. De Leeuw, and N. Benešová (2013), *European air quality maps of PM and ozone for 2011 and their uncertainty*, Bilthoven, Netherlands.

Horálek, J., P. de Smet, P. Kurfürst, F. De Leeuw, and N. Benešová (2014), *European air quality maps of PM and ozone for 2010 and their uncertainty*, ETC/ACM Technical Paper 2014/4.

Horálek, J., P. de Smet, P. Kurfürst, F. de Leeuw, and N. Benešová (2015), European air quality maps of PM and ozone for 2012 and their uncertainty: ETC/ACM Technical Paper 2014/4, , (January), 1–75.

Howe, J. (2006), The Rise of Crowdsourcing, *Wired Mag.*, *14*(6), 1–5, doi:10.1086/599595.

Irwin, A. (1995), *Citizen science: a study of people, expertise, and sustainable development*, Routledge.

Isaaks, E. H., and R. M. Srivastava (1989), *Applied geostatistics*, Oxford University Press, New York.

Josse, B., P. Simon, and V. H. Peuch (2004), Radon global simulations with the multiscale chemistry and transport model MOCAGE, *Tellus, Ser. B Chem. Phys. Meteorol.*, *56*(4), 339–356, doi:10.1111/j.1600-0889.2004.00112.x.

Journel, A. G., and C. J. Huijbregts (2003), *Mining geostatistics*, Blackburn Press.

Kalnay, E. (2003), *Atmospheric Modeling, Data Assimilation and Predictability*, Cambridge University Press, Cambridge, UK.

Kitanidis, P. K. (1997), *Introduction to Geostatistics: Applications in Hydrogeology*, Cambridge University Press.

Lahoz, W., and Q. Errera (2010), Constituent Assimilation, in *Data Assimilation*, edited by W. A. Lahoz, B. Khattatov, and R. Ménard, pp. 449–490, Springer Berlin Heidelberg, Berlin, Heidelberg.

Lahoz, W., B. Khattatov, and R. Menard (Eds.) (2010), *Data Assimilation*, Springer, Berlin, Heidelberg.

Lahoz, W. A., and P. Schneider (2014), Data assimilation: making sense of Earth Observation, *Front. Environ. Sci.*, *2*(16), 1–28, doi:10.3389/fenvs.2014.00016.

van Leeuwen, P. J. (2009), Particle Filtering in Geophysical Systems, *Mon. Weather Rev.*, *137*(12), 4089–4114, doi:10.1175/2009MWR2835.1.

Marécal, V. et al. (2015), A regional air quality forecasting system over Europe: The MACC-II daily ensemble production, *Geosci. Model Dev.*, *8*(9), 2777–2813, doi:10.5194/gmd-8-2777-2015.

De Nazelle, A., E. Seto, D. Donaire-Gonzalez, M. Mendez, J. Matamala, M. J. Nieuwenhuijsen, and M. Jerrett (2013), Improving estimates of air pollution exposure through ubiquitous sensing technologies, *Environ. Pollut.*, *176*, 92–99, doi:10.1016/j.envpol.2012.12.032.

Overeem, a., J. C. R. Robinson, H. Leijnse, G. J. Steeneveld, B. K. P. Horn, and R. Uijlenhoet (2013), Crowdsourcing urban air temperatures from smartphone batterytemperatures, *Geophys. Res. Lett.*, *40*(15), 4081–4085, doi:10.1002/grl.50786.

Pebesma, E. J. (2004), Multivariable geostatistics in S: The gstat package, *Comput. Geosci.*, *30*(7), 683–691, doi:10.1016/j.cageo.2004.03.012.

R Core Team (2016), R: A language and environment for statistical computing,

Rodgers, C. D. (2000), *Inverse Methods for Atmospheric Sounding*, World Scientific Publishing.

Rosner, H. (2013), Data on Wings, *Sci. Am.*, *308*(2), 68–73, doi:10.1038/scientificamerican0213-68.

Sarma, D. D. (2009), *Geostatistics with Applications in Earth Sciences*, Springer Science & Business Media, Dordrecht, The Netherlands.

Schaap, M., R. M. a. Timmermans, M. Roemer, G. a. C. Boersen, P. J. H. Builtjes, F. J. Sauter, G. J. M. Velders, and J. P. Beck (2008), The LOTOS EUROS model: description, validation and latest developments, *Int. J. Environ. Pollut.*, *32*(2), 270, doi:10.1504/IJEP.2008.017106.

Schmidt, H., C. Derognat, R. Vautard, and M. Beekmann (2001), A comparison of simulated and observed ozone mixing ratios for the summer of 1998 in Western Europe, *Atmos. Environ.*, *35*, 6277–6297.

Schneider, P., N. Castell, M. Vogt, W. Lahoz, and A. Bartonova (2017), Mapping urban air quality in near real-time using observations from low-cost sensors and model information, *Environ. Int.*, *accepted*.

Shanley, L. A., R. Burns, Z. Bastian, and E. S. Robson (2013), Tweeting Up a Storm The Promise and Perils of Crisis

mapping, *Photogramm. Eng. Remote Sens.*, (October 2013), 865–879.

Simpson, D., H. Fagerli, J. Jonson, S. Tsyro, and P. Wind (2003), *Transboundary Acidification, Eutrophication and Ground Level Ozone in Europe - Part I - Unified EMEP Model Description*, Oslo, Norway.

Slørdal, L. H., S.-E. Walker, and S. Solberg (2003), *The Urban Air Dispersion Model EPISODE applied in AirQUIS 2003 - Technical Description*, Kjeller, Norway.

De Smet, P., J. Horálek, M. Conková, P. Kurfürst, F. De Leeuw, and B. Denby (2010), *European air quality maps of ozone and PM10 for 2008 and their uncertainty analysis*, Bilthoven, Netherlands.

Sofiev, M., M. Galperin, and E. Genikhovich (2008), A Construction and Evaluation of Eulerian Dynamic Core for the Air Quality and Emergency Modelling System SILAM, in *Air Pollution Modeling and Its Application XIX*, pp. 699–701, Springer Netherlands, Dordrecht.

Spiegelhalter, D., M. Pearson, and I. Short (2011), Visualizing Uncertainty About the Future, *Science (80-.).*, *333*(6048), 1393–1400, doi:10.1126/science.1191181.

Wackernagel, H. (2003), *Multivariate Geostatistics*, Springer Berlin Heidelberg.

Webster, R., and M. A. Oliver (2007), *Geostatistics for Environmental Scientists*, John Wiley & Sons.