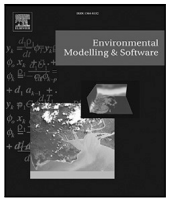
Environmental Modelling and Software 155 (2022) 105465

Contents lists available at ScienceDirect

Environmental Modelling and Software

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



Position Paper

Data fusion system for monitoring water quality: Application to chlorophyll-a in Baltic sea coast

M. Gunia [a,*], M. Laine [b], O. Malve [c], K. Kallio [c], M. Kervinen [c], S. Anttila [c], N. Kotamäki [c], E. Siivola [c], J. Kettunen [c], T. Kauranne [d]

[a] Lappeenranta University of Technology, Yliopistonkatu 34, 53850 Lappeenranta, Finland
[b] Finnish Meteorological Institute, Erik Palménin aukio 1, 00560 Helsinki, Finland
[c] Finnish Environment Institute, Latokartanonkaari 11, 00790 Helsinki, Finland
[d] Arbonaut Ltd., Kaislakatu 2, 80140 Joensuu, Finland

ARTICLE INFO

Keywords:
Water quality
Coastal
Data fusion
Data assimilation
Spatio-temporal interpolation

ABSTRACT

We present an operational system for multi-sensor data fusion implemented at the Finnish Environment Institute. The system uses Ensemble Kalman filter and smoother algorithms, which are often used for probabilistic analysis of multi-sensor data. Uncertainty and spatial and temporal correlations present in the available observation data are accounted for to obtain accurate and realistic results. To test the data fusion system, daily chlorophyll-a concentration has been modelled across northern shoreline of Gulf of Finland over the period of August 1st – October 31st 2011. Chlorophyll-a data from routine monitoring stations, ferrybox measurements, and data derived from Medium Resolution Imaging Spectrometer (MERIS) instrument on board the ENVISAT satellite has been used as input. The data fusion system demonstrates the use of existing and well-known Ensemble Kalman filtering and smoothing methods for improving water quality monitoring programs and for ensuring compliance with ecological standards.

1. Introduction

We describe a data fusion system (DFS) for water quality monitoring implemented at the Finnish Environment Institute SYKE. The goal of the system is to harmonize information from various data sources and to provide an estimate of water quality without data gaps, i.e. also at locations and times where observations are not available. To obtain accurate and realistic results, it is necessary to account for uncertainty in the observational data and exploit spatial and temporal correlations known to be present in the system. The uncertainty of the final estimates is quantified to better understand the limitations of the data fusion products and to help in designing better data collection strategies in the future. The data fusion products are available in the form of raster maps that can be directly visualized and published. Corresponding numerical data can also be queried interactively and exported from the system for further processing. The presented data fusion system uses Kalman filter and smoother algorithms, which are often used to analyse spatio-temporal variation of multi-sensor data. DFS is implemented as a general-purpose data fusion platform and is not limited to particular physical quantities. For the development and testing, chlorophyll-a (Chl-a) concentration and turbidity were used as the two primary water quality indicators to focus on. The main objective of the case study was to provide daily spatial Chl-a estimations that are useful in the ecological classification according to EU Water Framework directive.

Water quality observations from Finnish coastal waters have been collected regularly since the 1960s. Monitoring is typically based on laboratory analysis of water samples, automatic sampling from commercial ships, automatic fluorometric measurements from commercial ships and buoys, and satellite image processing. This gives us an increasing amount of heterogeneous environmental data with varying accuracy, precision, temporal frequency and regional coverage. Water analysis in laboratory is usually very accurate, but represents limited spatial and temporal domain. Observations derived from satellite data, on the other hand, are less accurate but cover much larger area and can be collected continuously with observations available daily or every few days. Data captured by satellites however often suffers from significant gaps due to the presence of clouds. Fluorometric measurements are obtained with the frequency of seconds or minutes and may cover a whole ship route or a single buoy location. To complicate matters further, the physical phenomenon of interest is often observed indirectly through proxies and indicators that are easier or more cost effective to obtain, compared to direct measurements.

* Corresponding author.
E-mail address: martin.gunia@student.lut.fi (M. Gunia).

https://doi.org/10.1016/j.envsoft.2022.105465
Received 21 February 2022; Received in revised form 24 May 2022; Accepted 11 July 2022
Available online 18 July 2022

Water quality of the Baltic Sea and Finnish coastal areas is continuously monitored for ecological classification and management according to the European Water Framework Directive (WFD, EU (2000)), EU Marine Strategy Framework (MSFD, EU (2008)) and the HELCOM Baltic Sea Action Plan (HELCOM, 2007). Granting of environmental permits for industrial and municipal waste water treatment plants, or any other loading operations, are based on ecological classification and therefore require rigorous monitoring and assessment of environmental status and impacts. Due to the precautionary principle of EU environmental law (Kriebel et al., 2001), an operation will not be permitted if there exists a risk of ecological deterioration of the receiving water body. As a result, monitoring needs to be precise and extensive to allow effective classification and permitting.

Combining environmental data from different sources and assimilating them to models of various complexity has been the subject of many methodological and application papers. We mention some of them here. Crow (2003) introduced data assimilation system of surface L-band brightness temperature (TB) observations via the ensemble Kalman filter (EnKF), to correct for the impact of poorly sampled rainfall on land surface model predictions of root-zone soil moisture and surface energy fluxes within the U.S. Southern Great Plains. Pulliainen et al. (2004) applied assimilation methods to combine ship-borne and satellite data of Chl-a in the Baltic Sea and to assess the spatial characteristics of water quality in the Baltic Sea, northern Europe. The technique is based on Bayes' theorem and considers spatial accuracy characteristics of both the transect and satellite data. Pan et al. (2008) implemented an integrated data assimilation system over the Red-Arkansas river basin to estimate regional scale terrestrial water cycle driven by multiple satellite remote sensing data. Mo et al. (2008) designed sequential data assimilation with an ensemble Kalman filter to optimize key parameters of the Boreal Ecosystem Productivity Simulator (BEPS) model, taking into account errors in the input, parameters and observation. A number of parameters were adjusted through data assimilation with a time step of one day. Chang and Latif (2010) modelled the behaviour of contaminants in a subsurface flow using two-dimensional transport model with advection and dispersion as the deterministic model. Stroud et al. (2010) studied space–time development of suspended sediment fields in lake Michigan using satellite data and ensemble Kalman methods. Melet et al. (2012) explored the potential use of glider data assimilation to control some properties of the ocean state estimation, such as thermohaline water circulation misfits in the Solomon Sea due to an erroneous tidal-mixing parametrization. The glider data was used to correct the model through a data assimilation scheme. Mourre and Chiggiato (2014) compared the ability of post-processing 3-D super-ensemble (3DSE) and conventional Ensemble Kalman filter (EnKF) approach integrating models and data to forecast the Ligurian Sea regional oceanographic conditions in the short-term range (0–72 h) when constrained by a common observation data set. Revilla-Romero et al. (2016) employed data assimilation techniques in hydrological forecasting to improve estimates of initial conditions and to update incorrect model states with observational data. Wang et al. (2019) presented a operational system for catchment-scale water quality management and monitoring and demonstrated its use in two test locations including Singapore's coastal waters and freshwater bodies. The system integrates input data from real-time sensors, measurements and models, a dynamic model of the catchment hydrology and data assimilation scheme to correct the model outputs with available observations. Fang et al. (2019) used space–time Kriging approach to estimate the dynamics of harmful algal blooms using Chl-a concentration measurements in Western Lake Eerie. The Kriging approach has been complemented with Bayesian information criterion for selection of explanatory variables to avoid model over-fitting and conditional simulation has been used to obtain probabilistic estimates. Qian et al. (2021) studied Chl-a and other indicators in Western Lake Eerie for modelling dynamics of cyanobacterial toxin concentrations. The authors developed hierarchical Bayesian framework for forecasting the toxin concentrations over time and demonstrated its applicability for short-term risk assessments. Chen et al. (2019) used data assimilation and the Ensemble Kalman filter to correct model forecasts of cyanobacterial biomass in Lake Taihu, China, using in-situ measurements and observations derived from remote sensing data. Recently, Chen et al. (2021) proposed data fusion method based on Bayesian inference principles similar to the ones used in this work and demonstrated its applicability to the estimation of Chl-a concentration in Lake Taihu using the in-situ and remote sensing observations. Other statistical and artificial intelligence methods have been used by Fasbender et al. (2008), Doña et al. (2015), Mouazen et al. (2014), and Chang et al. (2014). For a general reference to statistical methods on spatio-temporal data, we refer to Cressie and Wikle (2011), a general reference to ensemble Kalman filter methods in data assimilation is Evensen (2009). For an application of ensemble methods to high-dimensional models, similar to those used here, see Katzfuss et al. (2020). Alternatives to ensemble methods are different reduced rank methods such as those used by Zammit-Mangion et al. (2018) and Ma and Kang (2020).

1.1. System design

The data fusion system has been designed to streamline the entire modelling process, starting from downloading of the observation data and their harmonization, followed by the data fusion computation and subsequent storage and visualization of the final data sets. Its interface is aimed at expert users, who can conveniently carry out data fusion, validation and calibration, estimate the status of water bodies, design operational services and optimize the use of measurement resources. The overall design of DFS is shown in Fig. 1. Observation data is read into the system from spatial databases and open data services for in-situ and remote sensing data, operated by SYKE. An important aspect of the design is to allow automation of various workflows such as near real-time fusion of new observations as they become available. For this reason, the process has been separated into distinct steps that are implemented as stand-alone subroutines or commands. The processing steps include (1) model variable definition, (2) definition of model domain and computational grid, (3) data download and harmonization, (4) data fusion calculation and (5) data export and visualization. These "building blocks" can be executed manually, scheduled to run periodically, or scripted to implement more complex workflows. The data fusion inputs and results are stored in central database from where they can be queried in external GIS applications or published via a web portal.

Majority of the system's functionality is implemented in Python language and is designed to be easily extensible. Where available, open source software components have been used. The computational core is implemented as a general-purpose library called EnDAS (Ensemble Data Assimilation System) and its source code is available under an open source license as well (Gunia, 2018). For implementation details see Section 2, Auxiliary material can be found in Appendix A. Below, we define the components of the system design.

Model variable

Model variable is the quantity of interest for which data fusion is carried out, such as the Chl-a concentration or turbidity. Each model variable has a declared unit of measure and interpretation (in case of Chl-a, for example, the variable is assumed to correspond to the average Chl-a concentration in the top 5 metres of the water column). Model variables supported by the system are currently predefined and cannot be dynamically added by users. New variables can however be declared in the database by the system's administrator. Necessary transformations of the incoming observations are carried out automatically during the data download and harmonization step.

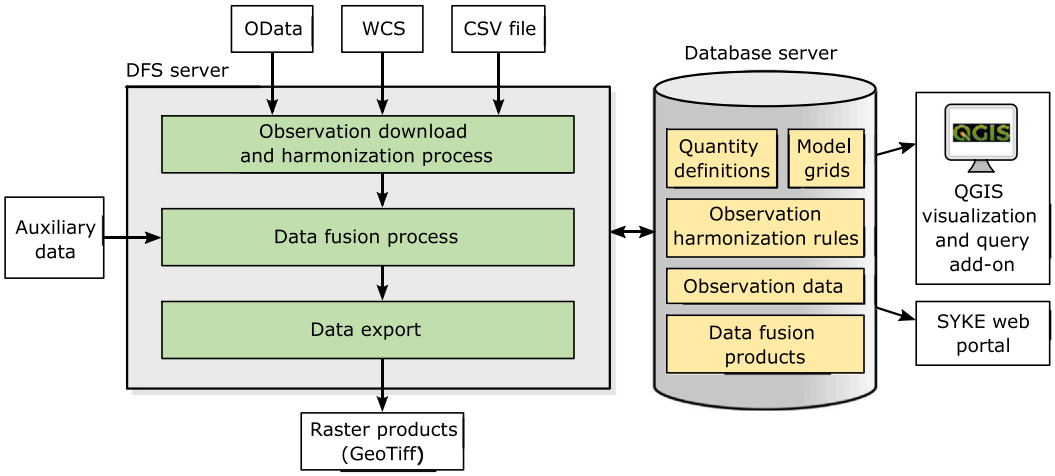


Fig. 1. Main components and information flow in the data fusion system as implemented at the Finnish Environment Institute SYKE.

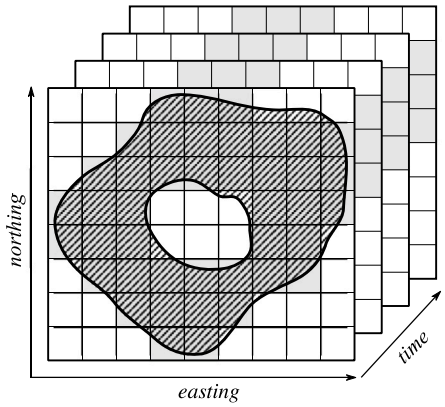


Fig. 2. Sparse model grid defined by three-dimensional array with two spatial and one temporal dimension. A hypothetical polygonal area defining the geographic extent of the model domain is indicated by diagonal hatching and only grid cells shown with grey background are included in the final array.

Model domain

Model domain defines the geographic extent of the model and is given as polygonal geometry by the user. The domain is spatially discretized using regular grid with two spatial and one temporal dimension. While the resolution of the spatial dimensions (cell size) can be given by the user, the temporal resolution and thus the frequency at which observations are fused is fixed to one day. The grid may be sparse and only grid cells which overlap with the domain geometry are actually stored in memory and included in computations, as shown in Fig. 2. Each stored cell then represents the value of a model variable, or multiple model variables for multivariate analysis, spatially integrated over the area of the cell.

While the underlying phenomena (such as water circulation and algae dynamics) are fundamentally three-dimensional, water quality indicators are typically concerned with the top of the water column. Consequently, observations are mainly available for the water surface or a shallow section of the water column. The two-dimensional depth-integrated approach is therefore sufficient for current DFS applications and requires significantly less storage capacity and computational power, compared to a fully three-dimensional model.

Data download and harmonization

Data download and harmonization refers to the download of observation data from configured data sources and subsequent pre-processing of the data to be suitable for data fusion. The downloaded data includes the actual measured values and metadata such as the acquisition date and time, uncertainty, unit of measure, measurement coordinate and station code. The data harmonization step consists of coordinate system transformation, spatial and temporal interpolation of the measured values and uncertainty to the model grid and unit conversion from the unit declared by the data source to the unit used by DFS. Additional transformations and corrections can also be applied, based on the observation type. These include correction to a common time-of-day (i.e. if observations are taken at different time than expected by DFS), depth correction or simple normalization. The system currently implements Open Data Protocol (OData) for discrete point measurement data, Web Coverage Service (WCS) for raster data sets, and Comma-Separated-Values (CSV) for offline point data. As a result of the data download and harmonization step, all available observations and their uncertainties are stored in the database and can be used directly by the data fusion process without the need for further pre-processing.

Data fusion

Data fusion is the process of combining information from observations to provide complete estimate of the model variables at all grid cells. The data fusion products include point estimates of the model variables and their uncertainty. Additional information such as the observation data sets used for the fusion, time and date of the modelling and data fusion settings used is also stored for later retrieval. The data fusion method is described in more detail in Section 2.

Data export and visualization

Data visualization capabilities are provided as a plug-in for the QGIS open-source geographic information system (QGIS Development Team, 2021). The plug-in is easy to install and can be used for browsing and querying of data stored in the DFS database. Additionally, data fusion products can be exported in GeoTiff raster format to enable further use and post-processing in external software.

1.2. Data fusion as a state estimation problem

The goal of data fusion is to integrate observations from multiple sources to produce more complete and accurate estimate than provided by each of the data sources alone. In a geophysical context, data is typically obtained from a scattered network of measurement devices and with varying sampling frequency. We therefore wish to perform statistically consistent interpolation of the data both in time and space, so that both the measurement uncertainty and spatial and temporal correlations in the data are accounted for. If the behaviour of the

observed physical system can be described by means a mathematical model, this additional information can be utilized, and we commonly refer to the resulting estimation procedure as *data assimilation*. Data assimilation has been initially developed for numerical weather prediction but has found its way into many other scientific and engineering disciplines such as GPS navigation, medical imaging or optimal control. In the context of the data fusion system, the terms "data fusion" and "data assimilation" overlap considerably and are many times used interchangeably. It should also be noted that although the currently implemented model (see Section 2.2) has no prediction power, DFS can be extended with more sophisticated dynamical models in the future; possible candidates are discussed in Section 4. The rest of this section provides brief summary of data assimilation theory that is relevant to the implemented system. It is not intended to be exhaustive and we refer the reader to Asch et al. (2016) or Evensen (2009) for more comprehensive introduction. Details of the implementation are provided in Section 2. In statistical terminology, the data fusion process can be seen as dynamical spatio-temporal data analysis, where an important aspect is the modelling of spatio-temporal correlations of the system processes, which are the key for realistic uncertainty quantification of the data fusion products. See Cressie and Wikle (2011) for a general reference to the statistical analyses.

The mathematical process model and the observations are two main components of any data assimilation system. The process model describes our understanding of the system's dynamics and its governing principles. Given an initial state of the system, the model can also be used to reason about which future states are more likely than others and to rule out states that are in contradiction with the underlying physical laws. Observations, on the other hand, are available at various times throughout the assimilation time frame and provide evidence of the real dynamics. We generally assume that observations are noisy, scarce and that the process of interest is often not observed directly. Even satellite remote sensing observations do not provide full coverage of the processes under study. Physical models can be used to constrain the solution of data assimilation analysis. The statistical approach assumes certain spatial and temporal correlations between the states. Observations, or the modelled states, that are close to each other in space and time are assumed, on average, to resemble each other more that observations that are further away. Because the initial state of the system is rarely well known and the mathematical model is an incomplete description of reality, model predictions will eventually deviate from the true state. Data assimilation aims at estimating the model state by correcting the state estimate proportionally to how much we trust the model and the observational evidence.

The main computational challenge in implementing efficient data fusion algorithms comes from the size of the modelled problem and consequently the amount of computer memory that is needed to store the system state (model variables for all grid cells) and to represent and manipulate the model error. To address this, DFS implements two widely used classes of algorithms. An exact Kalman Smoother (Kalman, 1960) algorithm can be used for model states of up to 10,000 elements and is suitable for example for small to medium-size lakes.

For large model grids, DFS implements ensemble-based Kalman Smoother (Evensen, 2009) where the system state and error is represented by a collection (ensemble) of possible realizations of the dynamical system. The number of realizations is typically chosen to be rather small while still being able to represent the majority of the model error. Furthermore, the ensemble algorithm is localized so that computations are only performed on a subset of the data at a time. This is motivated by the fact that the correlation between variables in any two grid cells decreases with distance and observations that are too far away are therefore assumed not to be of influence. This reduces sampling errors inherent to all ensemble-based approaches and allows much larger model grids to be processed. The size of the local window can be adjusted to balance the result quality and processing speed. Theoretical details of the implemented computational methods are explained in Appendix A.

2. Implementation overview

2.1. State space representation

The system design is based on the concept of a state space model, which can be written as two equations:

$$\begin{aligned}\mathbf{x}_k &= \mathcal{M}_k(\mathbf{x}_{k-1}, \boldsymbol{\theta}) + \boldsymbol{\eta}_k \\ \mathbf{y}_k &= \mathcal{H}_k(\mathbf{x}_k, \boldsymbol{\theta}) + \boldsymbol{\epsilon}_k.\end{aligned} \tag{1}$$

It describes the evolution of the multi dimensional process state $\mathbf{x}_k$ in time, with time index k. Observation vector $\mathbf{y}_k$ contains all available observations at time k and depends only on the current state. Both the evolution of the state and the observations contain uncertainties which are modelled by stochastic components $\boldsymbol{\eta}$ and $\boldsymbol{\epsilon}$. The other elements of the equation are explained below and we refer to Appendix A.1 for more technical details.

2.2. Evolution model

The role of the evolution model $\mathcal{M}$ in Eq. (1) is to propagate state of the system forward in time. DFS currently implements a random walk model with a drift, i.e. a simple linear model that converges to an a-priori "background" state over time. The model is given by

$$\mathcal{M}_k(\mathbf{x}|\boldsymbol{\mu}_\mathbf{b}, \alpha) = \alpha\left(\mathbf{x} - \boldsymbol{\mu}_\mathbf{b}\right) + \boldsymbol{\mu}_\mathbf{b}, \tag{2}$$

where $\boldsymbol{\mu}_\mathbf{b}$ is the background state vector and $0 < \alpha \leq 1$ is a dimensionless scalar factor controlling the rate of convergence. The background state is assumed to be constant for all grid cells and only the background mean value therefore needs to be chosen by the user. Different model variables can have different mean values. The use of the drift towards a mean state is motivated by the fact that events characterized by Chl-a peaks are often transient in nature and their typical time span is known from previous monitoring efforts. Thus, the data fusion system is instructed to return to the average state should no observations be available for a longer time period. The drift can be disabled by setting $\alpha = 1$, the model is then reduced to $\mathcal{M}_k(\mathbf{x}|\boldsymbol{\mu}_\mathbf{b}, \alpha) = \mathbf{x}$. In spite of the apparent simplicity, this model formulation can be used in a wide range of situations, where the dynamics of the processes cannot be directly modelled and the target is to augment the missing data spatially and temporally with realistic uncertainties. Here the actual modelling of the systems spatial correlations is based on the definition of the model error term $\boldsymbol{\eta}_k$ in the state-space Eq. (1) and, in the Kalman filtering context, the covariance matrix $\mathbf{Q}$ (see Section 2.4 for information on its construction). As such, the forecasting power of the current model is limited. However, the same formulation can be readily extend to more physics based models, for example in the form of a discretized advection–diffusion model (see, e.g. Stroud et al., 2010).

2.3. Observation model

The role of the observation model $\mathcal{H}$ is to describe the relationship between measurements in the observation vector $\mathbf{y}_k$ and the state vector $\mathbf{x}_k$. DFS currently assumes that data in the input data sources correspond directly to the modelled variables or, if not, any necessary transformations have already been performed as part of the data harmonization step (see Section 1.1). This leaves the observation model implementation rather trivial and the only required functionality is the spatial interpolation of measured values to the model grid for point observations (such as the monitoring station observations). Because gridded observations are already interpolated to the model grid during data harmonization, no additional interpolation is necessary.

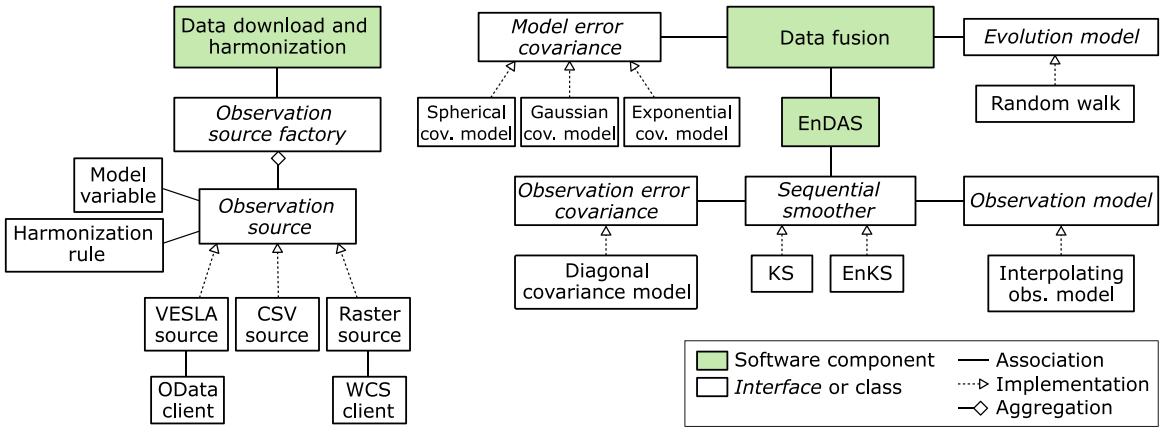


Fig. 3. UML class diagram of the main DFS extension points. Brief explanation of the interfaces is given in the main text and in Appendix.

2.4. Representation of uncertainty

In the state space framework, uncertainty is expressed by means of model and observation error terms $\boldsymbol{\eta}_k$ and $\boldsymbol{\epsilon}_k$. The model error accounts for the inability of the model to fully describe the real physical system, due to the lack of knowledge of the underlying governing principles, mathematical simplifications or errors introduced by numerical representation. Because the currently implemented evolution model has very limited forecasting power, the model error must account for all spatial and non-spatial uncertainty of the estimate. The Kalman filter and smoother assumes that errors are Gaussian and error terms $\boldsymbol{\eta}_k$ and $\boldsymbol{\epsilon}_k$ are therefore expressed by means of model and observation error covariance matrices $\mathbf{Q}_k$ and $\mathbf{R}_k$, respectively. In the full-rank Kalman smoother, the covariance matrices are full-rank and are included in the forecast and analysis step equations directly. In the ensemble filtering and smoothing context, on the other hand, manipulation of explicit covariance matrices is avoided and the effect of model error is implemented by perturbing the ensemble with random realizations drawn from $\mathcal{N}(\mathbf{0}, \mathbf{Q}_k)$. In both cases, the covariance matrix $\mathbf{Q}_k$ must be defined, either explicitly or implicitly.

In DFS, the model error covariance is assumed to be isotropic and can therefore be described as a function of distance $C(h) : \mathbb{R} \rightarrow \mathbb{R}$ such that $\mathbf{Q}_{i,j} = C(|i - j|)$. Here the time step index k has been omitted for clarity, although both $\mathbf{Q}$ and $\mathbf{R}$ can vary in time. Spatial variability of a random field is more commonly expressed in terms of a variogram $2\gamma(h) = \mathrm{Var}[X(i) - X(i + h)]$ rather than a covariance function; the variogram is related to the covariance function via the relationship $2\gamma(h) = C(0) - C(h)$. DFS implements several popular variogram models including the exponential, Gaussian and spherical models (see, e.g. Chiles and Delfiner, 2012; Schabenberger and Gotway, 2005, for the equations). A common assumption for sparsely scattered data is to assume observation errors to be independent between individual measurements and this assumption is also used by DFS. This simplifies the analysis scheme significantly because $\mathbf{R}$ is a diagonal matrix given by $\mathbf{R} = \mathbf{I}\sigma_{\mathrm{R}}^{2\mathrm{T}}$. The diagonal elements σ_{R}^2 are the observation variances, summarized in Section 3.2. It should be noted that the assumption of independence can be expected to hold sufficiently well for data that was collected by different instruments, such as the in-situ station measurement data, or data from different field campaigns. The assumption may be problematic for the satellite based data, as it is likely that errors in the neighbouring cell estimates derived from a single satellite scene are correlated. Correlated observation errors are discussed further in Section 4.

2.5. Software implementation

The DFS commands have been designed to be easy to modify an extend via well-defined software interfaces and protocols, which are presented in Fig. 3. Support for new observation data sources can be added by implementing the "observation source" interface. The interface is responsible both for the download of data and its harmonization because the harmonization procedure is specific to the data source. DFS implements interface to the open data service VESLA, which is part of the Environmental Information System of the Finnish Environment Administration. The data service contains results of physio-chemical measurements carried out by regional environment centres as well as private companies and water protection associations. The data is accessed through the Open Data Protocol. Additionally, gridded observations can be downloaded through the WCS (Web Coverage Service) interface and off-line point observation data can be read from CSV files. Multiple aspects of the data fusion algorithm can be customized as well. The software library EnDAS offers unified sequential smoothing API and several algorithms are implemented by the library. The API is non-intrusive (for ensemble methods) and does not require any interaction with the dynamic model. Therefore, the evolution model is considered a "black box" by the data assimilation scheme. DFS provides an interface for the evolution model to allow customization without the need to make changes in the assimilation algorithms. The observation model, which defines the relationship between observed values and modelled fields, can also be replaced by providing new implementation. Lastly, the model and observation errors are implemented via "covariance operator" interfaces provided by EnDAS. Covariance operators are abstract representations of covariance matrices that are typically defined implicitly in a lower-dimensional subspace. This way the usually prohibitive storage requirements of full-rank matrices are avoided and only the effects of the implicit matrices on data are computed.

The data fusion system is implemented in Python and runs on the Intel Distribution for Python (Intel Corporation, 2021). The Intel Python Distribution is a high-performance alternative to the reference Python implementation for computationally-intensive tasks. In addition, performance-critical parts of DFS for which the overhead of pure Python would be unacceptable are written using Numba (Lam et al., 2015). Numba is a just-in-time compiler for Python that transparently generates C code from the Python source, which is then compiled to machine code before it is executed. Unlike static compilers such as Cython, Numba does not rely on custom extensions to the Python language and infers efficient C code through introspection at runtime. Numerical arrays and linear algebra routines are provided by the NumPy and SciPy Python packages, respectively. Both packages are internally written in C and C++ and provide convenient Python interface. The Intel Distribution for Python comes with optimized NumPy, SciPy and Numba packages. NumPy and SciPy code is linked with the Intel Math Kernel Library (MKL), highly-optimized and scalable implementation of BLAS (Basic Linear Algebra Subprograms) and LAPACK (Linear Algebra PACkage) routines for multi-core CPUs. The DFS server runs Windows Server 2016 Standard operating system, although the implementation

Fig. 4. Typical work-flow in the DFS system.

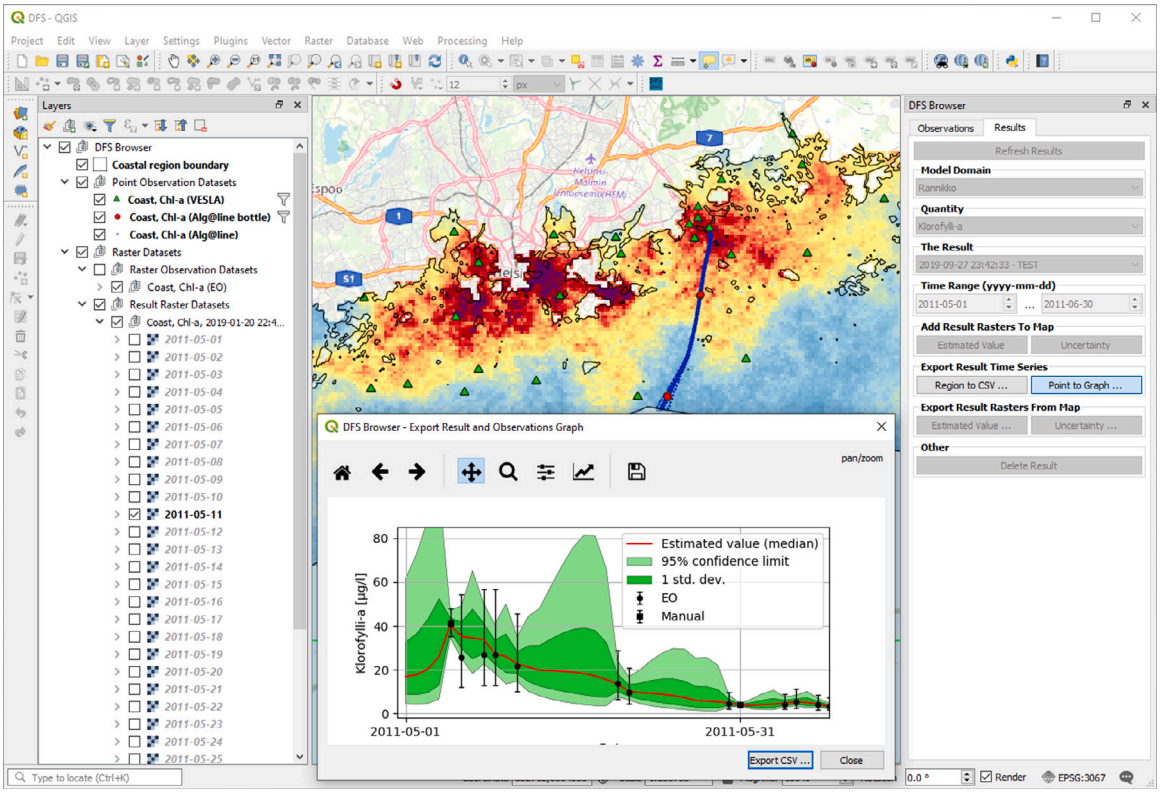


Fig. 5. DFS data visualization in QGIS 3 allows observations, fusion estimates and their uncertainties to be displayed on the map canvas and on an interactive time-series chart and to be exported in CSV format.

is portable and can be deployed, with minor modifications, on Unix-like platforms as well. For the database server, PostgreSQL version 9.5.13 is used, with the PostGIS spatial extension for handling vector and raster (gridded) data.

2.6. Processing tools and data visualization

The data fusion tools module consists of a set of stand-alone Python scripts to execute the full data fusion workflow, consisting of model domain creation, data import, data fusion calculation as well as result raster export in specified points of time or as time series in specified spatial points. Fig. 4 illustrates the usual DFS user workflow. In the beginning, user defines the model domain as the spatial area (polygon) for the analysis. User can use a model domain which already exists in the system or can add a new one to the database. When the model domain is available in the database, the next step is to run data harmonization. In data harmonization, the measurement data is read for the model domain and harmonized, i.e. converted to the internal coordinate system of the database, adjusted for time and depth, etc., and then saved to the database.

After the model domain and the observation data are processed and saved to the database, the data fusion can be run. When the data fusion is ready, the results are saved to the database. The data fusion results, as well as source data, can be viewed in QGIS. The DFS browser plugin for QGIS provides an easy to use user interface for the selection of the data from the data base. Also, point wise time series and individual raster maps can be exported with DFS Browser. The access to the observations, fusion estimates and their uncertainties through the DFS plug-in in QGIS 3 is handy to the end users. Any of them can be plotted on map canvas and on an interactive time-series chart for any location by clicking on the map. In addition, they can be exported in CSV format for further analysis. The data DFS visualization plug-in for QGIS is shown in Fig. 5.

3. Case study: Chl-a concentration in the Baltic coastal area

To demonstrate the use of the system, Chl-a concentration in the Baltic coastal area has been studied over one summer from April 1st till October 31st, 2011. In-situ observations from observing stations and measurement systems installed on board commercial ships (ferrybox) as well as observations derived from satellite-based remote sensing material were used. Within the modelled period there were 289 daily observations available from the monitoring stations, 3602 observations from the ferrybox instruments and 26 satellite images including millions of observation records. Due to the masking of clouds, the spatial coverage of the satellite data varied between 10 and 100 %.

3.1. Study area

Approximately 100 km long and 20 km wide stretch of the northern coastline of Gulf of Finland was selected for testing of the data fusion system. The area spans between the Porkkala peninsula and the

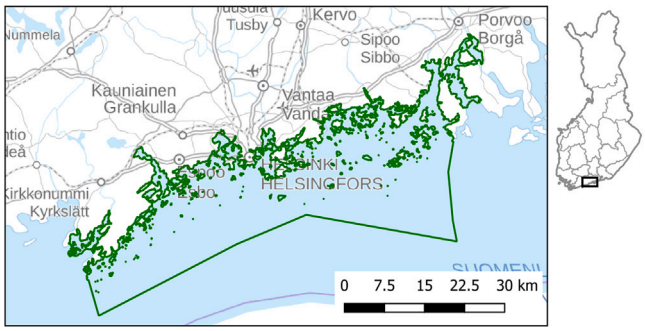


Fig. 6. Porkkala-Porvoo test area in southern Finland, shown in green. Background map from the National land survey of Finland.

archipelago of Porvoo, Finland, and includes the Helsinki metropolitan area. This area is divided into several water bodies for monitoring and management according to the Water Framework Directive. The outer water bodies define the border of the data fusion area towards the central Gulf of Finland. The area is relatively shallow and has low salinity between 0.2 and 5.8 ‰ at the surface and 0.3–8.5 ‰ near the bottom. The average water temperature is close to 0 °C in winter. In summer, it is 15–17 °C at the surface and 2–3 °C at the bottom. The area can freeze from late November to late April. The coast is abundant in small bays and skerries, but includes only a few large bays and peninsulas (such as the Porkkala peninsula). The main nutrient loading sources are municipalities and agriculture, resulting in eutrophication and occasional cyanobacteria blooms in July–August. Water quality in the area is regularly monitored by various methods and water management and pollution control measures are under way. The overview of the area is presented in Fig. 6.

3.2. Available data

All available in situ observations of Chl-a were collected from the study area. These include manual water samples taken at routine monitoring stations and ferrybox measurements collected under the Alg@line project (Seppälä et al., 2007). The station sample data is available via the VESLA data service and accessed by DFS over the Open Data Protocol. VESLA is a part of the Environmental Information System of the Finnish Environment Administration (SYKE, 2020) and includes physio-chemical measurement results of national and regional monitoring, carried out by regional environment centres, as well as local statutory monitoring results conducted by private companies and water protection associations. The Alg@line data, collected on board the m/s Finnmaid Ro-Ro/passenger ship, includes sensor measurements of Chl-a fluorescence. These were converted to Chl-a concentration using water samples taken during the cruise that were analysed in laboratory. The Alg@line water sampling measurements of Chl-a were also input to the DFS. Typically, 24 samples are taken on the return trip from Germany to Helsinki. The in-situ sampling data of the routine monitoring stations were obtained from the VESLA OData service.

Satellite data with continuous spatial coverage were also available. We used the data from Medium Resolution Imaging Spectrometer (MERIS) instrument, that was onboard the ENVISAT satellite and operational between 2002–2012. Version of the L1B satellite data was 3rd data reprocessing with MERIS Ground Segment (MEGS) Processor Version 8.0. Dataset geolocation was further refined with the AMORGOS (Accurate MERIS Ortho-Rectified Geo-location Operational Software) tool, version 3.0. The biophysical parameters, such as Chl-a, were derived from the L1B dataset using a neural network-based processor FUB/WeW WATER (Schroeder et al., 2007b,a), available in the SNAP software.

The individual data sources represent different sections of the water column. Water samples for Chl-a at the routine monitoring stations were taken as a composite sample from 0 m down to two times the Secchi depth transparency. Secchi depth in the study area usually varies between 1 and 5 m, mainly depending on the distance to land and river mouths, and on the time of the year. The depth that the satellite observations represent depends on the concentrations of the colour-producing substances (typically total suspended matter, Chl-a and humic substances), and is commonly estimated by the attenuation depth. About 90 % of the back-scattered light from a water column to the atmosphere comes from the surface layer down to the attenuation depth. In the southern coast of Finland, attenuation depth is about 50 % of the Secchi depth, estimated by the method described in Kallio et al. (2015). The nominal sampling depth of the Alg@line data is the depth of the water intake (5 m). However, the ship hull mixes the water before the sample is taken and the Alg@line measurements are thus considered to represent the layer of 0–5 m. We assume in the data fusion that the water column is fully mixed from surface down to the deepest depth of the mentioned measurement depths. The observation data is summarized in Table 1.

In addition to the measured values, DFS also requires information about uncertainty of the input data to produce correct results. Measurement uncertainty of manual water samples taken at routine monitoring stations is laboratory-specific and in most cases an estimate is available. The uncertainty reported by laboratories operating in Finland usually takes into account both the random and systematic error factor and is given as a standard deviation or its multiple. In the case of Chl-a concentration, uncertainty is given as relative standard deviation and normal distribution is assumed for the measurement error. The reported uncertainty is stored as metadata in the VESLA database and can therefore be directly retrieved by DFS. Where not available, the average uncertainty of water samples collected along the southern Finnish coast has been used as a default value. Uncertainty of data derived from satellite images was quantified in an earlier validation studies conducted in the Finnish coastal waters (Attila et al., 2018) and is likewise assumed to be Gaussian. Uncertainty of the Alg@line water samples was obtained from the Marine laboratory of the Finnish Environment Institute. For the flow-through sensor measurements, the average uncertainty of continuous buoy measurements conducted in the years 2013–2014 (Kallio et al., 2015) was used. Measurement errors used in this work are summarized in Table 2.

In order for the Chl-a concentration to remain positive and to account for the error heterogeneity (error being proportional to the estimated concentration), data fusion has been carried out using logarithms of the values and using relative standard deviations as uncertainties. More details on performing data fusion in logarithmic scale are given in Appendix A.4.

3.3. Spatial auto-correlation

Statistical variogram analysis was performed on the available MERIS data for Chl-a. Exponential, Gaussian, Matérn and spherical variogram models (see, e.g. Chiles and Delfiner, 2012; Schabenberger and Gotway, 2005) were fitted for all days for which data was available between June and August 2011. The spherical variogram model resulted in the lowest average fitting error and has been selected to represent the spatial dependency of Chl-a variance. Typical correlation length in June and August was approximately 12 km. In July, the correlation length varied more than in June and August and ranged between 8 and 15 km. July is often characterized by patchy cyanobacterial blooms and the lower end of the fitted range, i.e. 8 km, was selected to be a suitable representative value based on expert judgement. Parameters of the fitted variogram model are summarized in Table 3.

Table 1
Summary of data used in the data fusion of Porkkala-Porvoo coastal area.

Data type	Source	Spatial resolution	Temporal resolution	Depth
Satellite sensor	MERIS on board ENVISAT	300 × 300 m	Approximately 2 days, only cloud-free data can be utilized	Depends on water quality, see text
Ferrybox, sensor measurement	Alg@line (m/s Finnmaid)	Approximately 200 m (20 s interval)	Measured with three day interval	5 m (nominal), see text
Ferrybox, water sampling	Alg@line (m/s Finnmaid)	Three sampling locations in the study area	Samples were taken 2–4 times per month	5 m (nominal), see text
Manual water sampling at monitoring stations	Statutory monitoring, monitoring by environmental authorities and cities	Discrete point	1–12 samples between April and October, depending on station	Composite sample from 0 m down to two times the Secchi depth

Table 2
Chl-a measurement error of data sources listed in Table 1. The uncertainty is given as relative standard deviation (RSD) and normal distribution is assumed.

Data source	RSD
Satellite sensor	38%
Ferrybox, sensor measurement	20%
Ferrybox, water sampling	15%
Monitoring stations, manual sampling	Varying, default 8%

Table 3
Fitted variogram parameters for Chl-a in the Porkkala-Porvoo area during June, July and August 2011. Spherical variogram model is assumed and the fitting was performed on logarithmic scale. The number of days for which variograms were fitted is indicated by *n*.

Month	Nugget (log μg/l)	Sill (log μg/l)	Range (km)	*n*	Fitting error (log μg/l)
June	5.7×10^{-3}	4.3×10^{-2}	12.8	4	2.6×10^{-6}
July	4.2×10^{-3}	2.7×10^{-2}	8.0	8	1.2×10^{-6}
August	6.2×10^{-3}	3.8×10^{-2}	12.6	4	7.3×10^{-7}

3.4. Other parameters

The cell size of the model grid cell has been set to 100 × 100 metres, resulting in state space containing approximately 130,000 elements. The ensemble size has been set to 200 members to represent the uncertainty in the system and to provide good trade-off between the quality of the result and execution time. The model error is described by a spherical variogram function with the non-spatial variance term set to 0.2 (applied on logarithmic scale) and the correlation length set to 8 km in July and 12.7 km otherwise. The data assimilation time window (i.e. the smoother lag) has been set to 10 days based on expert judgement. The smoother lag is used by ensemble filters and smoothers to limit the influence of observations that are too far away in time from the analysis. The value of 10 days has been judged large enough because peaks in Chl-a concentrations are typically much shorter. The initial state of the system, i.e. its mean and error covariance has been calculated automatically by the data fusion system. The initial mean is calculated as the mean over all observations over an interval of 3 days centred at the starting day of the data fusion run. The initial error is calculated by taking the 95-percentile of the relative observation error on each day within the same interval and taking their average. The parameters used in the data fusion run are summarized in Table 4.

3.5. Case study results

Reconstructed Chl-a concentration and its error estimate for a selected day (July 7th, 2011) is shown in Fig. 7. Because of the 10-day temporal window used, the Chl-a concentration for that day was estimated from MERIS instrument data, the Alg@line ferry data and data from measurements stations available between June 28th and July 17th (i.e. within 10 days from July 7th). The Chla-observations derived from the MERIS instrument for July 7th, the Alg@line ferry route and locations of stations from which data was available within the 10-day window are shown in the top section of Fig. 7. The estimated Chl-a concentration is presented in the middle section and the 95% confidence interval of the estimate is shown in the bottom section of Fig. 7.

Fig. 7 demonstrates the ability of the system to reconstruct the Chl-a concentration over the entire model domain in the presence of relatively large data gaps. The spatial variability and small-scale variation of the estimated field is well captured, mostly due to the high-resolution observations from the MERIS instrument. The estimated confidence interval reflects both the data sparsity and the fact that errors are assumed to be proportional to the estimated value. Therefore, the error is largest in areas for which observations are not available for a prolonged period, but which have high estimated Chl-a concentration. The high Chl-a concentration may be either an estimate obtained earlier or an effect of the influence of more distant observations. In the current simulations, the most severe data gaps are found in the areas near the shoreline, where no MERIS data coverage was present. These areas are however very significant from the water quality management perspective. The confidence interval output can be conveniently presented in the form of spatial maps which can be used in making decisions on further data acquisition work.

The estimated daily Chl-a concentrations are also shown in Fig. 8 for four selected locations and utilizing all observation data (left column) and a subset only containing the satellite (EO) observations. The four selected locations are shown in Fig. 7, top. Point 1 is located in the centre of the archipelago area, point 2 in the open sea area and points 3 and 4 in the inner archipelago area. Chl-a concentration was monitored frequently at points 1 and 2 and infrequently at point 3. Satellite observations were also less frequent in the location of point 3 and were missing at point 4 altogether. Satellite observations were however available few hundred metres away from point 4. In general, the confidence interval of the interpolated concentrations was larger with EO-data only, compared to confidence interval with all data. This was also the case in the inner archipelago area at points 3 and 4 with infrequent or fully missing satellite observations. We see that the interpolated concentrations are close to in-situ observations when only satellite data are available nearby. Further outside of the satellite data coverage, the estimated concentrations are inaccurate. In 2011, Chl-a followed seasonal pattern typical in the Gulf of Finland: spring peak occurring in late April or early May and late summer peaks occurring in July and August. In spring, phytoplankton is typically dominated by Dinoflagellates and Diatoms, in late summer by Cyanobacteria. Timing and intensity of late summer peaks depend on the weather conditions (low wind speed and high temperature favour Cyanobacteria accumulations). During the declining phase of Chl-a peak in the beginning of July 2011, there was a period for which satellite observations were not available, making the estimation of the duration of the peak somewhat uncertain. Manual samples were available in points 1 and 4 during that time, which assisted the detection of the decline near the measurement locations. On wider scale, however, manual samples did not make a significant difference. After July, notable Chl-a peaks were not observed. Based on the novel multi-source observations at the northern coast of the Gulf of Finland, spatial and temporal distribution of Chl-a appeared rather transient and multifaceted. The simulated system is therefore highly dynamic and is characterized by short and

Table 4
Data fusion model run parameters.

Parameter	Value	Description
Grid size	100 m	Cell-size for data harmonization and data fusion.
Ensemble size	200	Size of ensemble in EnKS.
Model error (log scale)	0.2	Additive model error term applied to the ensemble at each time step. This corresponds to the sigma- term of the covariance function.
Correlation length	8 km in July, 12.7 km otherwise	See Section 3.3
Smoother lag	10	To limit the influence of future observations over time (value in time steps)
Smoother forgetting factor	0.7	The forgetting factor is used in conjunction with the lag to reduce the influence of future observations during the smoothing stage.
Covariance inflation	1.02	Slight inflation to compensate for ensemble sampling errors.
Initialization	Automatic	The initial mean and error covariance is estimated by the data fusion system from observation data.

often spatially-localized peaks in Chl-a concentrations. The availability of frequent observations with sufficient spatial coverage is crucial for properly capturing these events. Calm and warm weather, which is favourable for phytoplankton growth, is likely to coincide with cloud-less sky. This allows satellite observations to be obtained and thus aims estimation of the Chl-a peaks.

The DFS system makes it easy to experiment with various options for the model setup and to study available observation sources and the information content they provide for global analysis on the whole study area. The possibility to fill the gaps in the observations, both temporal and spatial, makes it easy to produce aggregated water quality assessments which include estimate of the uncertainty coming from the interpolation process. Estimation of Chl-a during periods with no observations could be improved by adding a hydrodynamic ecosystem model to the DFS.

To assess the sensitivity of the DFS results to observation errors, we adjusted the observation errors artificially and made simulations with 18% and 58% errors (addition to the default 38%) for EO data, and with 10% and 30% errors (in addition to the original 20%) for Alg@line sensor data. The estimates varied by a rather narrow margin and the sensitivity to observation errors was minor. As an example, the mean relative Chl-a errors over the period 1–31 July 2011 at point 2 were 25.2%, 28.4% and 30.2% for EO observation errors 18%, 38% and 58% (the corresponding estimated mean Chl-a concentrations were 4.55 μg/l, 4.87 μg/l and 4.54 μg/l, respectively). For the varying Alg@line observation errors 10%, 20% and 30%, the mean relative errors for the same time period and location were 27.9%, 28.4% and 29.1% (the corresponding mean Chl-a concentrations were 4.76 μg/l, 4.87 μg/l and 4.60 μg/l). It should however be noted that the effect of observation error in both satellite and Alg@line sensor data is likely underestimated due to the fact that in both data sets the observations and their errors are likely correlated (due to their close spatial proximity). These correlations are currently not handled by the system and may lead to underestimation of the error as discussed in Section 4. The results of the sensitivity runs are presented more in detail in Appendix C in the Auxiliary material.

4. Discussion and conclusion

We have presented an operational system for multi-sensor data fusion implemented at the Finnish Environment Institute. To test and evaluate the data fusion capabilities, daily Chl-a concentration has been modelled for a part of northern shoreline of the Gulf of Finland, including the Helsinki metropolitan area. The modelling has been performed for the period between April 1st and October 31st 2011, utilizing data collected from manual sampling stations, automatic flow-through measurements collected on-board commercial cruise vessels and data derived from satellite imagery. The application of the data fusion system to the monitoring of coastal area in the Gulf of Finland has shown feasibility and potential of the system for improving water quality monitoring and management.

The implemented data fusion methods allow processing of relatively large model domains using high-resolution satellite observations. The use of the exact, full-rank algorithm is limited to approximately 10,000 state variables due to the need to manipulate covariance matrices explicitly. This is sufficient for many lakes and does not require tuning of auxiliary parameters of the model. The ensemble-based algorithm, on the other hand, operates on a reduced-rank representation of the error covariance and thus requires significantly less computer memory. In addition, the algorithm is localized and the size of model grid is therefore only limited by the amount memory needed to store the grid cells, rather than a covariance matrix. This limit is however seldom met in practice and for practical applications the limit on the model grid is given by the time available to run the data fusion. The execution time of the results presented in Section 3 has been approximately 2 h using local window size of 3 × 3 cells. Increasing the local window to 5 × 5 cells lowers the execution time to approximately 1 h without noticeable loss of quality. The data fusion system runs on a dedicated server machine with Xeon E5-2667 processor (16 physical cores at 3.2 GHz) and 64 GB of RAM.

It is a general challenge in data fusion to obtain model parameters that give logical and realistic results. This is even more so with the ensemble-based algorithms that rely on tuning parameters to work well. The development of DFS has been motivated by the need for an operational system that streamlines processing of observations and their spatio-temporal interpolation and that allows for testing and experimentation with the available data sets. Thus, we have focused primarily on the implementation of the data fusion framework and tentative parametrization of the system has been attempted. Rigorous derivation of the needed parameters has been left out for future work.

The MERIS satellite data used in this work contains significant gaps and does not cover small bays. Therefore, the data fusion system cannot currently produce realistic Chl-a estimates in these areas. This can be improved in the future by using data from the Sentinel-2 Copernicus mission which has much improved spatial coverage. Also, the spatial resolution of the Sentinel-2 data is 10–20 m, compared to 300 m resolution of the MERIS data. The dynamic model currently implemented in DFS has no prediction power. The system can however be easily extended in the future with more sophisticated dynamical models. An example would be a model that takes known seasonal trends into account or an advection–diffusion model similar to the one used in Stroud et al. (2010). Known limitation of the current implementation is that observation errors are assumed to be independent and are thus conceptually represented by a diagonal covariance matrix. While this holds sufficiently well for observations from measurement stations and buoys, which are farther apart, the assumption may be problematic for observations derived from satellite data or fluorometer samples. In both cases, individual observations are close to each other and their observation errors are likely to be correlated. Treating such observations as independent leads to underestimation of uncertainty of the data fusion result due to observation errors partially cancelling each other out (this can be seen in Fig. 7). It should be noted that the data fusion algorithm can naturally handle correlated observation errors by

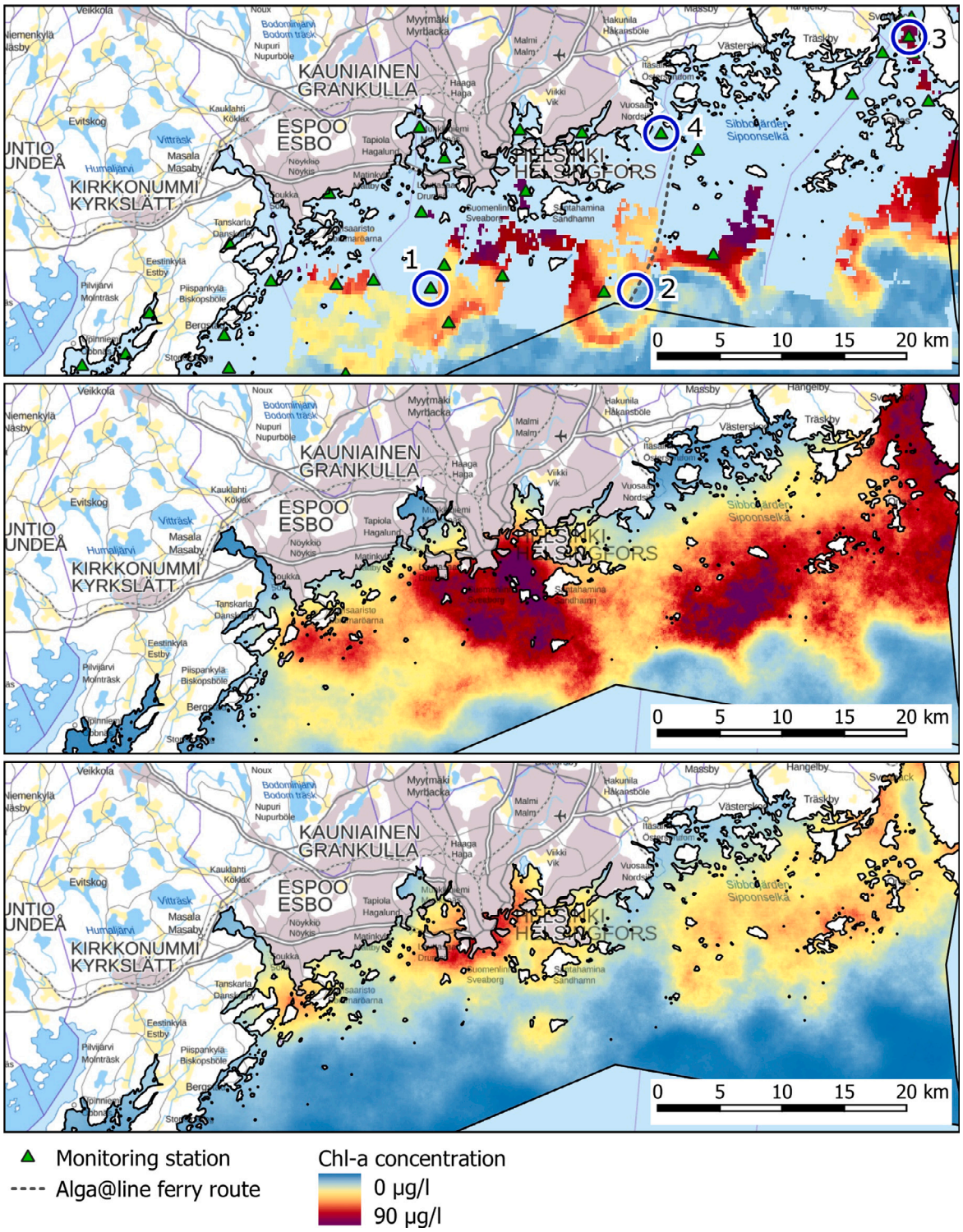


Fig. 7. Reconstructed Chl-a concentration on July 7th, 2011. The top figure shows the observed Chl-a concentration available from the MERIS data, the Alg@line ferry route and locations of measurement stations for which data was available between June 28th and July 17th (i.e. within 10 days from July 7th). The reconstructed Chl-a field is shown in the centre and its uncertainty in the bottom figure as the width of the 95 % confidence interval. Time-series from four monitoring points 1–4 are plotted in (Fig. 8). Background map from the National Land Survey of Finland.

choosing an appropriate observation error covariance. Because of the high dimension of the state space, the error covariance must however be expressed as a low-rank approximation of the real covariance and this has deliberately been left out for future work. Additionally, correlations between multiple sensors can be handled in the same way as correlations within the same sensor. Because data from one satellite sensor only has been used in this work, these correlations were also left out for further study. Additional future developments of the system include the use of co-variate information, such as water temperature, to improve the accuracy of the results and automatic estimation of model parameters from available data by the system.

If there is a risk for deterioration of ecological status of coastal area, precision of the data fusion results is focal for the monitoring and management of water quality in accordance with the EU Water

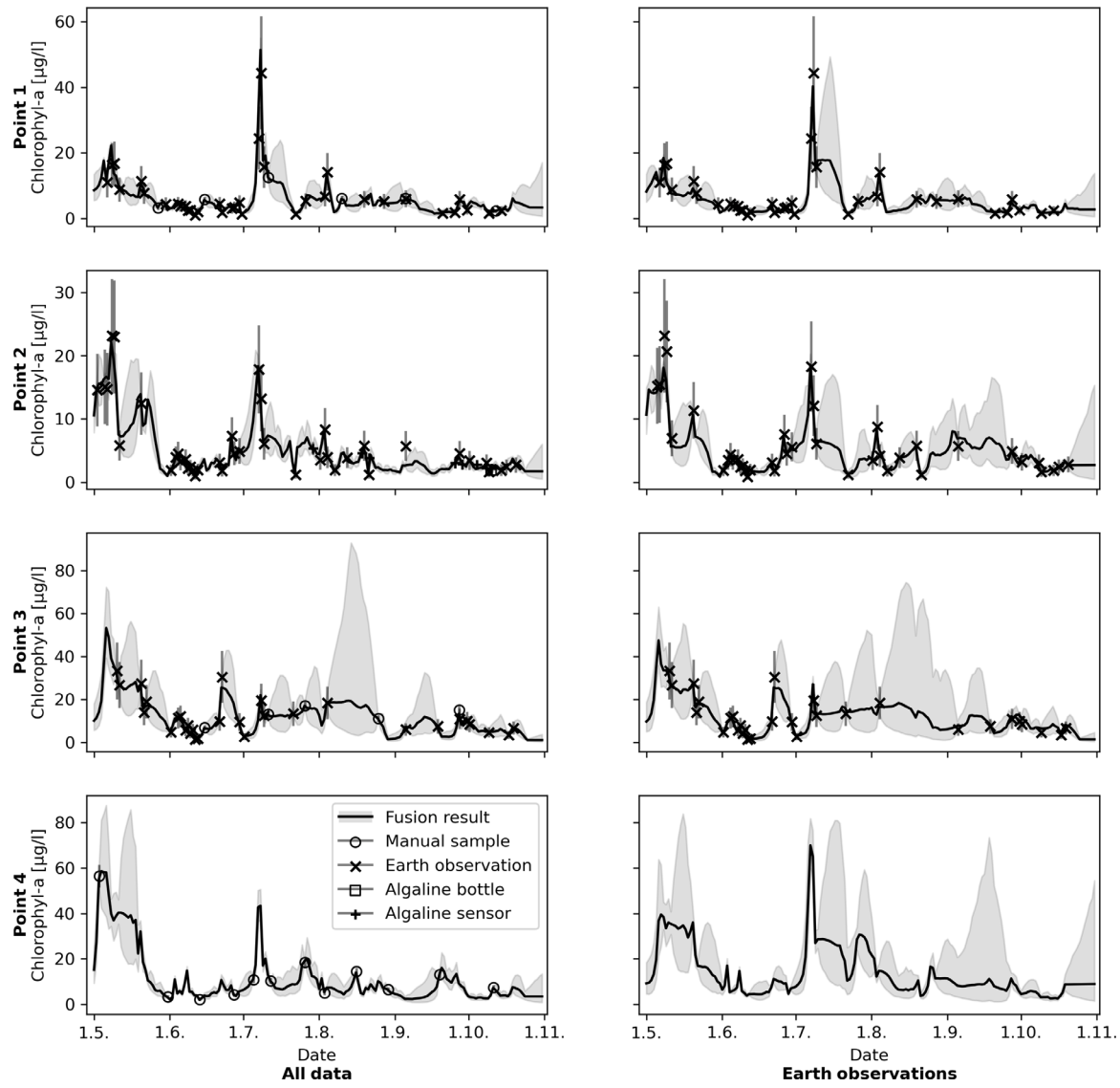


Fig. 8. Reconstructed Chl-a concentration for points 1–4 (see Fig. 7 for locations of the points). (Left) Estimates generated using all available observation data, (right) estimates using satellite observations only. The best Chl-a concentration estimate (µg/l) is shown with black line, uncertainty is shown in grey as the width of the 66 % confidence interval. Observed values are shown with varying symbols (see the legend) and error bars correspond to one standard deviation.

Framework Directive (EU, 2000). Moreover, environmental permit of a polluter may be rejected based on the precautionary principle of EU environmental law (Kriebel et al., 2001). On this account, monitoring programs need to be optimized, improved and extended to reduce error variances, confidence limits and the risk of non-compliance with ecological standards. The data fusion system makes it possible to take the full advantage of the available data sources to get more complete estimate of the water quality and ecological status. One possible application is ecological classification and management of coastal and inland waters according to the WFD. It can also be used for environmental monitoring and permitting of fish farms or any other polluter and to assess the impacts of these activities more precisely than by conventional methods.

Software availability

Software name: EnDAS
Year of first release: 2019
Operating systems: Windows, Linux, Mac
Programming languages: Python
Availability: https://github.com/martingu11/endas
License: MIT
Documentation: https://endas.readthedocs.io/en/latest

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported financially by the Strategic Research Council of the Academy of Finland (BLUEADAPT project, grant number 312650), the Finnish Ministry of the Environment and the Earth and Water Engineering Support Association (Maa- ja vesitekniikan tuki ry). Part of the work was also funded by the ADAFUME project of the Academy of Finland (project number 321890).

Auxiliary material

See Appendices A–C.

Appendix A. Data fusion algorithms

A.1. State space representation

Here we formally define the data assimilation system. Let $\{X_k\}$ be a stochastic process over a set of time steps $k, 0 \leq k \leq K$ which represents the dynamical system of interest and the state at time step k is denoted as X_k. The process is assumed to have the Markov property, so that future system states only depend on the history thru the current state, i.e. the future is independent of the past given the present. We further assume that $\{X_k\}$ is unobservable (i.e. its state is "hidden") but another process $\{Y_k\}$ exists that is observed and is dependent on $\{X_k\}$ in some known way. In other words, $\{X_k\}$ is a hidden Markov chain. The time step $k = 0$ corresponds to the initial system state and we will further assume that observations are available at one or more time steps $k, 1 \leq k \leq K$. The states of $\{X_k\}$ are numerically represented by a sequence of n-dimensional vectors of real numbers $\mathbf{x}_k \in \mathbb{R}^n$. Similarly, observations are represented by r_k-dimensional vectors $\mathbf{y}_k \in \mathbb{R}^{r_k}$ where r_k is the number of observations at time step k.

Given a mathematical model of the process $\mathcal{M} : \mathbb{R}^n \to \mathbb{R}^n$, the evolution of the system can be described by state space equations

$$\begin{aligned} \mathbf{x}_k &= \mathcal{M}_k(\mathbf{x}_{k-1}, \boldsymbol{\theta}) + \boldsymbol{\eta}_k \\ \mathbf{y}_k &= \mathcal{H}_k(\mathbf{x}_k, \boldsymbol{\theta}) + \boldsymbol{\epsilon}_k. \end{aligned} \tag{3}$$

The model $\mathcal{M}_k$ is often called the *evolution* or *dynamic model* and expresses the relationship between consecutive system states by means of a model prediction. $\mathcal{H}_k$ is called the *observation model* or observation operator and is a function $\mathcal{H}_k : \mathbb{R}^n \to \mathbb{R}^{r_k}$ that maps the state $\{X_k\}$ to observations $\{Y_k\}$. Vector $\boldsymbol{\theta}$ contains auxiliary model parameters, which typically do not depend on the time k. Finally, $\boldsymbol{\eta}_k$ and $\boldsymbol{\epsilon}_k$ are additive terms that account for the model and observation errors, respectively. Both are assumed to be random, independent from each other and independent in time. It should however be noted that although universally accepted, the assumption of independence is rarely fully true in reality. Data collected by the same instrument over a period time, for example, may have correlated errors on subsequent measurements.

The final step in completing the data assimilation scheme is to combine the evolution model with observations. In a probabilistic formulation, we aim to estimate the posterior distribution of the state space conditioned on collected observations: $p(\mathbf{x}_{0:k}|\mathbf{y}_{1:q})$, where $k \leq q \leq K$. Here we use the notation $\mathbf{x}_{a:b}$ to denote the sequence of state vectors $\mathbf{x}_{a:b} = \{\mathbf{x}_a, \mathbf{x}_{a+1}, \dots, \mathbf{x}_b\}$ and $\mathbf{y}_{a:b}$ to denote the sequence of observation vectors defined in an identical fashion. Depending on the choice of q, the following data assimilation schemes can be distinguished:

- $p(\mathbf{x}_{0:k}|\mathbf{y}_{1:k})$ – We wish to estimate system states up to and including the current step k, using all observations collected so far. This is an instance of so-called *Bayesian filtering*
- $p(\mathbf{x}_{0:k}|\mathbf{y}_{1:k+l})$ for $l > 0$ – We wish to estimate system states up to and including the current step k, using observations up to and including a future time step $k + l$. The parameter l is called the *lag* and the scheme is referred to as *fixed-lag Bayesian smoothing*
- $p(\mathbf{x}_{0:K}|\mathbf{y}_{1:K})$ – We wish to estimate system states over the entire data assimilation window, using all available observations. This is called *fixed-interval Bayesian smoothing*.

For the data fusion applications and dynamical spatio-temporal data analysis, the main tool is the smoother as it allows to combine all available information both in time and space. The joint probability densities mentioned above are usually not practical and the computationally simpler marginal distributions $p(\mathbf{x}_k|\mathbf{y}_{1:q})$ are used instead. The filtering posterior distribution $p(\mathbf{x}_k|\mathbf{y}_{1:k})$ can be obtained by applying Bayes' rule as

$$p(\mathbf{x}_k|\mathbf{y}_{1:k}) \propto p(\mathbf{y}_{1:k}|\mathbf{x}_k)p(\mathbf{x}_k|\mathbf{y}_{1:k-1}), \tag{4}$$

where $p(\mathbf{x}_k|\mathbf{y}_{1:k-1})$ is the *prior distribution* and describes the probability of $\mathbf{x}_k$ before the evidence in $\mathbf{y}_k$ is considered. $p(\mathbf{y}_{1:k}|\mathbf{x}_k)$ is the observation likelihood conditioned on $\mathbf{x}_k$ and therefore contains new information present in observations when contrasted with the current state estimate. This is also sometimes called the *innovation*. From Eq. (4) we can see that the filtering solution can be obtained as soon as an observation becomes available, allowing on-line data assimilation. The assumption of uncorrelated errors and the Markovian property of $\{X_k\}$ allows the posterior smoothing distribution to be calculated as a product

$$p(\mathbf{x}_{0:K}|\mathbf{y}_{1:K}) \propto p(\mathbf{x}_0) \prod_{k=1}^{K} p(\mathbf{y}_k|\mathbf{x}_k)p(\mathbf{x}_k|\mathbf{x}_{k-1}), \tag{5}$$

where $p(\mathbf{x}_0)$ is the distribution assigned to the initial state. As with the filtering solution, the smoothing solution for a given time step k, $p(\mathbf{x}_k|\mathbf{y}_{1:K})$, can be obtained by marginalization of (5). This is often desired because it avoids re-computation of the entire joint posterior distribution of $\mathbf{x}_{0:K}$ every time new observations are obtained. Furthermore, the fixed-lag scheme relies on the marginalized solutions to update state estimates for the last l steps so that the resulting data assimilation can still be considered on-line but with a fixed delay.

A.2. Kalman filter and smoother

The large dimension of state spaces encountered in geosciences usually prevents direct manipulation of the probability density functions in Eqs. (4) and (5) and we resort to approximations. The Kalman filter (KF, Kalman (1960)) is a popular and often used approach that efficiently solves the filtering and smoothing problems for *linear dynamical systems*. The prior and posterior distributions are approximated by Gaussians, which can be fully described by the first two statistical moments—the mean and covariance. The model and observation errors in Eq. (3) are thus assumed to be normally distributed:

$$\boldsymbol{\eta}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_k) \text{ and } \boldsymbol{\epsilon}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_k), \tag{6}$$

where $\mathbf{Q}_k$ and $\mathbf{R}_k$ are the model and observation error covariance matrices, respectively. The posterior density given by Eq. (5) then becomes a product of Gaussians and is therefore also Gaussian. The marginal filtering and smoothing posterior distributions at a time step k, $p(\mathbf{x}_k|\mathbf{y}_{1:k})$ and $p(\mathbf{x}_k|\mathbf{y}_{1:K})$, are then characterized by covariance matrices $\mathbf{P}_{k|1:k}$ and $\mathbf{P}_{k|1:K}$, respectively. The subscript notation $k|1:q$ is analogous to the notation used in Eqs. (4) and (5) and denotes the error covariance at time k so that information from observations up to and including the time step q has been incorporated.

Kalman filter is a sequential filtering algorithm for state estimation that consists of two alternating steps. First, the state vector $\mathbf{x}_{k-1}$ and the corresponding error covariance matrix $\mathbf{P}_{k-1|1:k-1}$ is propagated from time step $k-1$ to k by application of the dynamic model. This is called the *forecast step*. The forecast step is followed by assimilation of observations called the *update* or *analysis* step. When observations are introduced, the state estimate is corrected and the posterior covariance is reduced. For actual Kalman filter equations, we refer the reader to existing literature such as Evensen (2009) or Asch et al. (2016). Kalman smoother is a direct extension of the Kalman filter and several formulations exist. The most direct approach is to "augment" the state vector of the filter at any given time step with state variables from previous time steps, effectively implementing a fixed-lag Kalman smoother. Alternatively, fixed lag Kalman smoother can be formulated without the need to increase the size of the state vectors through retrospective updates (Cohn et al., 1995). In this approach, the state vectors $\mathbf{x}_j$ at $k - l \leq j < k$ are updated after the assimilation of observations into $\mathbf{x}_k$ at time t_k. Another well known approach is to first compute the Kalman filter solutions for all $k = \{1, \dots, K\}$ and then proceed with a backward updating pass for $k = \{K-1, \dots, 0\}$ in which information from observations used to correct state estimates at earlier

time steps. The forward–backward technique is known as the Rauch–Tung–Striebel smoother (Rauch et al., 1965). Its advantage is that each state vector is only updated twice and a fixed-interval solution over the entire window $k = 0 \dots K$ is obtained. However, the potentially very large covariance matrices $\mathbf{P}_{k-1|1:k-1}$ must be stored for all time steps during the filtering pass. This may be prohibitively expensive in terms of storage, even if the matrices are written to disk storage until they are needed again.

Kalman filter and smoother can be shown to be an optimal unbiased estimator of the posterior distributions if the system is linear and errors are Gaussian. Real processes are however seldom linear and non-linear generalizations, such as the extended Kalman filter (EKF) and Unscented Kalman filter (UKF, Julier et al. (1995)), have been developed. EKF relies on linearization of the dynamic and observation models $\mathcal{M}_k(x)$ and $\mathcal{H}_k(x)$ through first-order Taylor series expansion around x. UKF, on the other hand, aims to directly approximate the posterior distribution rather than approximating the process and observation models. The approximation is computed purely through evaluations of $\mathcal{M}_k(x)$ and $\mathcal{H}_k(x)$, model linearizations are therefore not necessary.

Creation of the linearized models, and their adjoints, required by EKF is a formidable task for complex models even with the help of automatic code differentiation tools. Another, more fundamental drawback is the need to manipulate full-rank covariance matrices (or sigma vectors in UKF) during the forecast and update steps. This makes Kalman filters, including EKF and UKF, unpractical for larger spatial domains such as those often encountered in geosciences where the size of the state vector may be $n \gg 10^5$. This has lead to the development of reduced methods for data assimilation.

A.3. Ensemble formulation

To overcome the large computational requirements of Kalman filters, an alternative Monte Carlo approach called ensemble Kalman Filter (EnKF) has been proposed by Evensen (1994). Instead of a single sequence system states $\mathbf{x}_k$ and covariances $\mathbf{P}$ for $k = 0, 1, \dots, K$, EnKF uses a collection or *ensemble* of states to approximate the prior and posterior distributions. The size of the ensemble is typically much smaller than the size of the state space, allowing ensemble Kalman filters to be used on very large systems.

Let the ensemble be defined by an $n \times N$ matrix

$$\mathbf{E}_k = \left[\mathbf{x}_k^{(1)}, \mathbf{x}_k^{(2)}, \dots, \mathbf{x}_k^{(N)}\right], \tag{7}$$

where N is the size of the ensemble and $\mathbf{x}^{(1)} \dots \mathbf{x}^{(N)}$ are the individual state vectors. We can then define a matrix of ensemble perturbations

$$\mathbf{E}'_k = \frac{1}{\sqrt{N-1}}\left(\mathbf{E}_k - \overline{\mathbf{E}}_k\right), \tag{8}$$

where $\overline{\mathbf{E}}_k = \mathbf{E}_k \mathbf{1}_N$ is a matrix holding the ensemble mean in each column and $\mathbf{1}_N$ is $N \times N$ matrix with each coefficient equal to $1/N$. The normalization factor $1/\sqrt{(N-1)}$ is chosen so that the full-rank error covariance $\mathbf{P}_k$ is replaced by sample covariance $\hat{\mathbf{P}}_k = \mathbf{E}'_k \mathbf{E}'^{\mathrm{T}}_k$. It should be noted that just like $\mathbf{P}_k$, $\hat{\mathbf{P}}_k$ is a $n \times n$ matrix and can therefore not be included in the filtering and smoothing equations explicitly. Instead, the product $\mathbf{E}_k \mathbf{E}'^{\mathrm{T}}_k$ is used directly and equations are rearranged so that the full matrix does not need to be computed.

The use of a finite and usually small ($m \ll n$) ensemble to represent the covariance matrix $\mathbf{P}$ comes with some drawbacks. The small sample size introduces errors into the estimated covariance, leading to non-zero correlations between state variables that in reality are physically unrelated. Because of these additional, spurious correlations, state variables that would normally be unaffected are corrected during the update step and their uncertainty therefore decreases. As a result, the state error covariance may become strongly underestimated over time, causing the filter to become overconfident about the model predictions. Observations become irrelevant and the filter generally diverges from the real system state. Filter divergence due to the underestimation of error covariance is a problem common to all ensemble Kalman filters and approaches such as covariance inflation and localization, have been developed to circumvent these issues. It is thanks to these practical "fixes" that ensemble Kalman filters have become very popular and successful tools for data assimilation. Covariance inflation and localization techniques implemented in DFS are described further in Appendix B.3.

A.4. Logarithmic scale

In the state space model presented in Section 1.2 the terms $\boldsymbol{\eta}_k$ and $\boldsymbol{\epsilon}_k$ represent absolute model and observation errors. In many situations, however, it is desirable to model the errors as relative and hence proportional to the measured or estimated value. A convenient way to achieve this is to perform data fusion in a transformed space

$$Z_k = \log\left(X_k\right) \tag{9}$$

and assimilate observations of $\log\left(Y_k\right)$ instead. Consequently, Z is normally distributed as $Z \sim \mathcal{N}\left(\mu_Z, \sigma_Z^2\right)$ and X becomes log-normally distributed according to $X \sim \log\mathcal{N}\left(\mu_X, \sigma_X^2\right)$ with the transformed mean and variance given by

$$\begin{aligned} \mu_Z &= \log\left(\frac{\mu_X^2}{\sqrt{\mu_X^2 + \sigma_X^2}}\right) \\ \sigma_Z^2 &= \log\left(1 + \frac{\sigma_X^2}{\mu_X^2}\right). \end{aligned} \tag{10}$$

In the equations above, the time indices k have been dropped to simplify the notation. Given the transformation above, the final point estimate $\mathbf{x}_k$ can be obtained as

$$\mathbf{x}_k = \mu_X = \exp\left(\mu_Z + \frac{\sigma_Z^2}{2}\right). \tag{11}$$

Appendix B. Implementation

B.1. Implemented Kalman smoothers

Because not all model grids need to be very large, DFS implements both the exact, full-rank Kalman smoother and an ensemble based Kalman smoother. The full-rank Kalman smoother is presented in Appendix A.2 and will not be described further. The ensemble Kalman smoother variant implemented in DFS is the Error Subspace Transform Kalman Filter (ESTKF) proposed in Nerger et al. (2012) and its smoother extension, (ESTKS, Nerger et al. (2014)). ESTKF belongs to the square root family of Kalman Filters that does not rely on random perturbations in the analysis step. Unlike traditional EnKF, ensemble-transform Kalman filters perform all algebra in a smaller space of the perturbations spanned by th ensemble members. The update scheme of ESTKF is very similar to that of the popular Ensemble Transform Kalman Filter (Bishop et al., 2001) but at a slightly lower computational cost.

B.2. Sampling strategy for EnKS

The Monte-Carlo approach of EnKS relies on random perturbations $\mathbf{A}' \in \mathbb{R}^{n \times N}$ drawn from $\mathcal{N}(\mathbf{0}, \mathbf{Q}_k)$ to implement the error term $\boldsymbol{\eta}_k$. One approach to generate the perturbations is to compute the square root of $\mathbf{Q}_k$, via Cholesky decomposition, and multiply it by a vector of independent random samples drawn from $\mathcal{N}(\mathbf{0}, \mathbf{I})$. This method is exact but computationally infeasible for large model grids due to the high computational cost of the Cholesky factorization. More efficient approach is to perform the sampling in Fourier space using the circulant embedding method (Wood and Chan, 1994), which is implemented in

DFS. It should be noted that the circulant embedding method requires the sampled grid to be regular (and dense), samples are therefore generated also for cells that are outside of the model boundary and not included in the model grid. These samples are currently discarded, which is inefficient if the model grid is very sparse. To remedy this, the block variant of the circulant embedding method (Park and Tretyakov, 2015) could be used to minimize the amount of discarded samples.

Because the size of the ensemble is typically much smaller than the size of the state space, effort has to be made to avoid unnecessary noise in the ensemble which can lead to unwanted correlations between state variables. The common technique for reducing the ensemble sampling noise is covariance localization, described in Appendix A.3. In DFS, however, the model error perturbations can also be a major source of noise because they are likely to dominate the system evolution, given the simplistic evolution model. The easiest way to improve the quality of the ensemble is to increase the its size. However, doing so leads to increased storage and computational cost of the data fusion algorithm, making large ensembles unpractical. Better approach is to keep the size of the ensemble fixed and aim to maximize its rank instead. One simple approach, based on the idea introduced in Pham (2001), can be implemented as follows: We first generate a larger, "augmented" ensemble of perturbations $\hat{\mathbf{A}}' \in \mathbb{R}^{n\times\omega N}$ sampled from $\mathcal{N}(\mathbf{0}, \mathbf{Q}_k)$ using the circulant embedding method. The size of the ensemble is ωN, with $\omega > 1$ being the "oversampling ratio". We compute the singular value decomposition

$$\hat{\mathbf{A}}' = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}}$$

and obtain the final ensemble $\mathbf{A}'$ by sampling along the N largest eigenvalues

$$\mathbf{A}' = \mathbf{U}\mathbf{\Sigma}_{1:N}\mathbf{\Theta}^{\mathrm{T}}.$$

Here $\mathbf{\Sigma}_{1:N}$ contains the first N eigenvalues of $\mathbf{\Sigma}$, multiplied by $\sqrt{N/(\omega N)}$, and $\mathbf{\Theta} \in \mathbb{R}^{N\times N}$ is a random orthogonal matrix. It should be noted that the augmented ensemble $\hat{\mathbf{A}}'$ and the singular value decomposition only need to be computed once. After that, only the random matrix $\mathbf{\Theta}$ needs to be computed to obtain the new sample $\mathbf{A}'$. The effect of the improved sampling scheme can be evaluated by looking at the singular values of the generated samples, as compared to those of the original sample. Fig. 9 demonstrates the improvement in the conditioning of the ensemble (by means of the ratio between the largest and smallest singular value) for a hypothetical model grid of size 300×300 cells and the spherical covariance function with a range of 30 cells. In this example, the oversampling ratio $\omega = 4$ provides a good compromise between improvement in quality and computational cost. Beyond $\omega = 6$, the improvement is minimal.

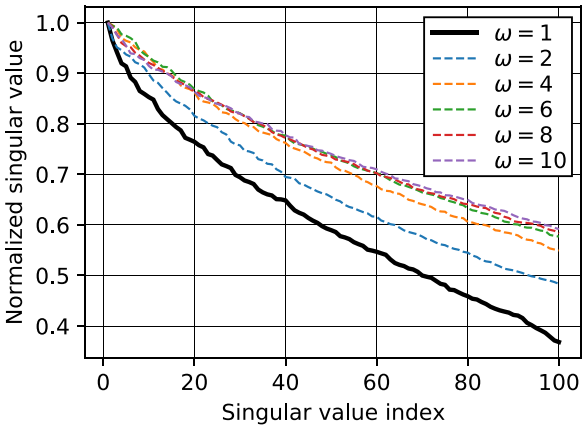


Fig. 9. Singular values of ensemble perturbations generated using the maximum-rank scheme. The singular value of the original ensemble are shown by think black line, dashed lines show singular values for $\omega \in \{2, 3, 4, 8 \text{ and } 10\}$. The singular values are normalized to the largest singular value.

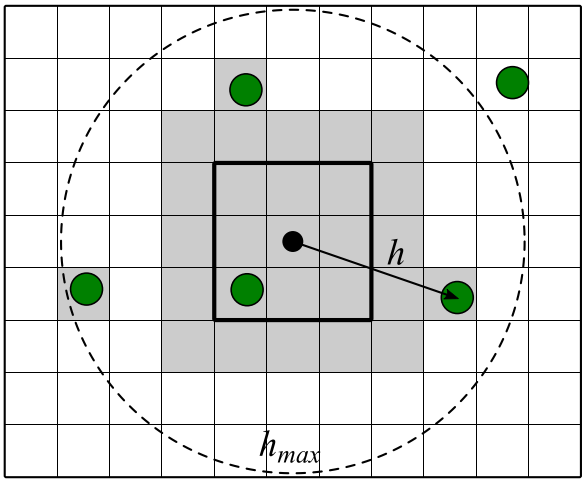


Fig. 10. Domain-localized update on a two-dimensional grid. One local analysis domain, 3×3 cells in size, is outlined in black and the locations of nearby observations are shown by green circles. Four observations are within the distance limit h_{max} and will be used for correcting the local state estimate. Cells whose state variables participate in the update are shown with dark shading. This also includes a one-cell padding around the local domain.

B.3. Covariance inflation and localization

Covariance inflation and localization are two techniques that are often used to stabilize ensemble Kalman filters, i.e. to avoid filter divergence due to overconfidence in the filter's performance. The purpose of covariance inflation is to account for underestimation of the true error covariance $\mathbf{P}$ by $\hat{\mathbf{P}}$ due to the limited size of the ensemble. While the forced inflation of the covariance does not have strong foundation in theory, it is nevertheless practical and easy to implement. To inflate the covariance $\hat{\mathbf{P}}$, matrix of ensemble anomalies $\mathbf{E}'$ from Eq. (8) is multiplied by a factor slightly larger than one. The factor may be a constant scalar value or may vary in time and space. Because of the ad-hoc nature of covariance inflation, the inflation is typically tuned for specific data assimilation problems by means of trial and error. However, methods that aim to estimate optimal inflation factors have also been proposed in Raanes et al. (2019), Miyoshi (2011) and Evensen (2009), to name a few. The approach suggested by Evensen (Section 15.3 2009) applies EnKF update equations directly to an auxiliary state to estimate the inflation factor. It has been tested with DFS but has not yielded consistent improvements and DFS therefore uses a constant inflation factor with a default value 1.02. Moreover, the use of covariance localization typically reduces the need for inflation and its extensive tuning.

Covariance localization aims to improve the estimated covariance $\hat{\mathbf{P}}$ by suppressing unwanted correlations, i.e. those that are assumed to be the product of the sampling error. In EnKS variants that operate in the state space, the localization effect can be achieved by regularizing the (implicit) covariance matrix $\hat{\mathbf{P}}$ through manipulation of the ensemble anomalies. However, ESTKS operates in the ensemble subspace and the direct regularization approach is not feasible. Instead, localized analysis is achieved by dividing the state vector into subsets called *local analysis domains*. Assimilation of observations is then carried for each local analysis domain separately, utilizing only observations within certain distance h_{max} from the domain. The distance of an observation from the local domain can also be used to reduce its influence so that far away observations contribute less to the local analysis solution. To do that, the observation uncertainty σ^2_{R} from Section 2.4 is additionally tapered by the covariance function $C(h)$ and becomes

$$\sigma^2_{\mathrm{R},i} = \sigma^2_{\mathrm{R},i}\left[\frac{C(h_i)}{C(0)}\right]^{-1} \tag{12}$$

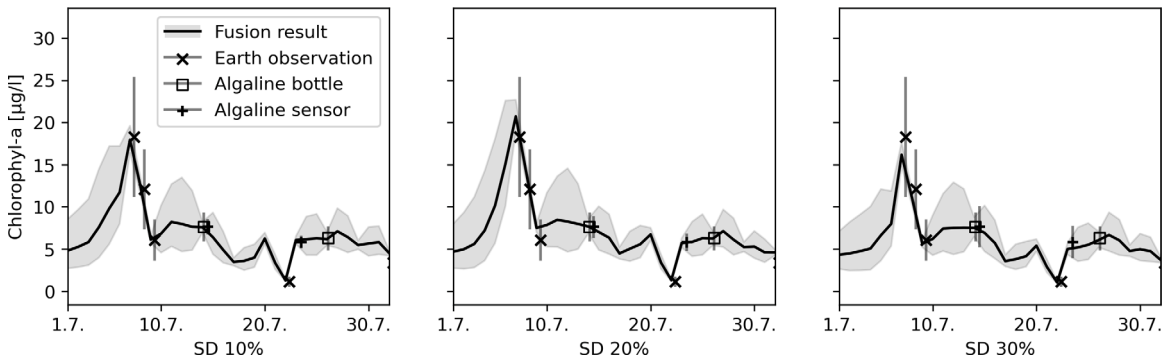


Fig. 11. Estimated Chl-a concentration in July for Point 2 with Alg@line observation error 10% (left), 20% (middle) and 30% (right). See Fig. 7 for location of the point.

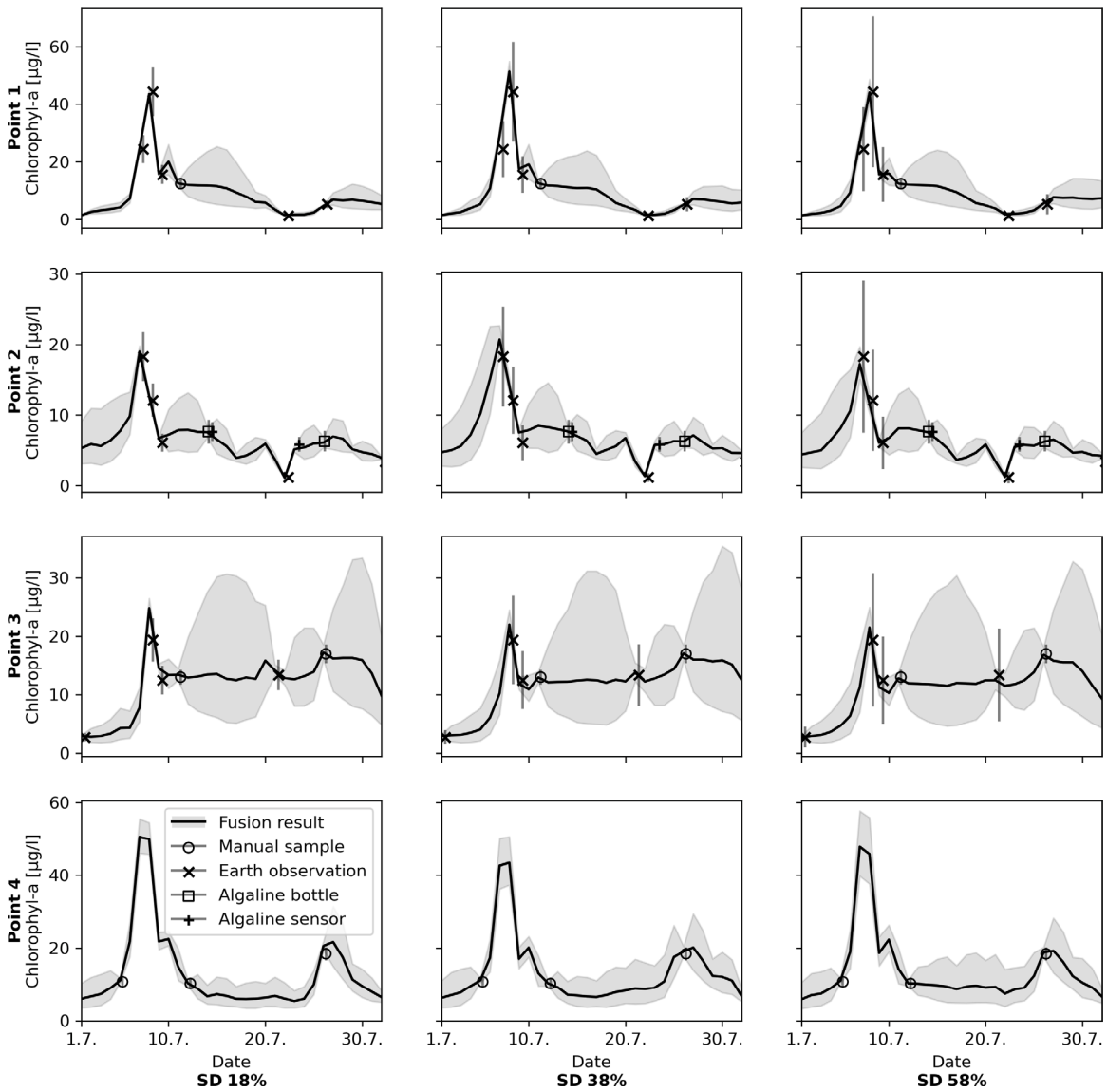


Fig. 12. Estimated Chl-a concentration in July for points 1–4 with satellite (EO) observation error 18% (left column), 38% (middle column) and 58% (right column). See Fig. 7 for locations of the points.

where h_i is the distance of the ith observation from the local domain and i runs over all observations in the local domain. The cutting distance h_{max} is chosen so that $C(h_{max})/C(0)$ is sufficiently small.

At minimum, each local domain may comprise of a single grid cell. The domains may however be larger for performance reasons and DFS uses square blocks of $r \times r$ grid cells for the local analysis, where

Table 5
Mean Chl-a concentration (μg/l) and the corresponding mean error (μg/l) in July for selected observation errors. See Fig. 7 for locations of the four points.

Location	Observation error	Mean Chl-a (μg/l)	Mean error (μg/l)	Mean of daily relative errors (%) (computed from daily estimates)
Point 1	EO 18%	5.48	1.62	33.4
	EO 38%	5.69	1.96	39.0
	EO 58%	5.50	2.12	41.7
Point 2	EO 18%	4.53	1.05	25.2
	EO 38%	4.87	1.32	28.4
	EO 58%	4.54	1.30	30.2
Point 3	EO 18%	12.8	8.23	58.1
	EO 38%	11.8	7.79	64.7
	EO 58%	11.5	7.51	66.0
Point 4	EO 18%	12.18	4.56	40.0
	EO 38%	10.98	4.38	45.4
	EO 58%	10.76	4.68	47.9
Point 2	Alg@line 10%	4.76	1.24	27.9
	Alg@line 20%	4.87	1.32	28.4
	Alg@line 30%	4.60	1.25	29.1

r is a tuning parameter that balances execution speed and quality. The default block size is $r = 3$ and local domains are also enlarged (padded) so that adjacent domains partly overlap. The overlap is used to smoothly blend ensembles from local domains back into the global ensemble to eliminate visible boundaries. The principle of local analysis in DFS is shown in Fig. 10.

B.4. Generation of initial ensemble

The data fusion system can estimate the initial state (the state vector mean and the covariance matrix) from available observations. A commonly approach used in data assimilation is to sample the initial ensemble from a long model run while maximizing the rank of the ensemble (Pham, 2001). This approach is however not currently feasible in DFS due to the trivial nature of the evolution model. Therefore, DFS attempts to estimate the initial state mean and variance from available observations by collecting available observations over a short interval $T = [t_{-N}, t_N]$, where $N = \min(l/3, 3)$ and l is the smoother lag or temporal auto-correlation length. This choice is made so that only observations that are believed to be reasonably close to the initial state, but at minimum ± 3 days are used. The initial state vector is then set to the arithmetic mean of the collected observation values, per model variable. Because the observational data is likely to be distributed unevenly in space, the background error is first calculated as the 95%-percentile of relative observation standard deviation of data points for each day in T. The daily errors are then averaged to yield the final initial state error standard deviation σ_{x_0}. Finally, the covariance matrix $\mathbf{P}_0$ needed for EKF is constructed analogously to the construction of the model error covariance matrix $\mathbf{Q}$ described in Section 2.4.

Appendix C. Observation error sensitivity

To assess the sensitivity of the data fusion estimates to the uncertainty in the observed data, the observation error has been artificially adjusted and additional data fusion simulations were run over the period of 1–31 July, 2011. The mean Chl-a concentration and the corresponding error of the data fusion runs has been recorded. The sensitivity runs were made with 18% and 58% relative error for the satellite (EO) data, and 10% and 30% relative error for the Alg@line data, in addition to the default observation errors used (38% for EO and 20% for Alg@line data). The mean Chl-a concentration and the corresponding mean errors for points 1–4 (see Fig. 7 for locations of the points) are shown in Table 5. The estimated Chl-a time series are also shown in Fig. 11 for the varying EO observation error and Fig. 12 for the varying Alg@line error.

The effect of varying observation errors on data fusion uncertainty is rather minor for both the satellite and Alg@line observations. The reason for this is the high number of observations available in both data sets that are in the vicinity of analysed points. Due to the assumption of error independence, the observation errors cancel each other out, leading to a seemingly precise estimate. The errors between Chl-a derived from individual satellite pixels and obtained from subsequent fluorometer samples are however likely to be correlated.

References

Asch, Mark, Bocquet, Marc, Nodet, Maëlle, 2016. Data Assimilation: Methods, Algorithms, and Applications. SIAM.

Attila, J., Kauppila, P., Alasalmi, H., Kallio, K., Keto, V., Bruun, E., Koponen, S., 2018. Applicability of earth observation chlorophyll-a data in assessment of water status via MERIS – with implications for the use of OLCI sensors. Remote Sens. Environ. 212, 273–287. http://dx.doi.org/10.1016/j.rse.2018.02.043.

Bishop, Craig, Etherton, Brian, Majumdar, Sharanya, 2001. Adaptive sampling with the ensemble transform Kalman filter. Part I: Theoretical aspects. Mon. Weather Rev. 129, http://dx.doi.org/10.1175/1520-0493(2001)129<0420:ASWTET>2.0.CO;2.

Chang, Shoou-Yuh, Latif, Sikdar Muhammad Istiuq, 2010. Extended Kalman filtering to improve the accuracy of a subsurface contaminant transport model. J. Environ. Eng. 136 (5), 466–474. http://dx.doi.org/10.1061/(ASCE)EE.1943-7870.0000179, URL https://ascelibrary.org/doi/abs/10.1061/%28ASCE%29EE.1943-7870.0000179.

Chang, Ni-Bin, Vannah, Benjamin W., Yang, Y. Jeffrey, Elovitz, Michael, 2014. Integrated data fusion and mining techniques for monitoring total organic carbon concentrations in a lake. Int. J. Remote Sens. 35 (3), 1064–1093. http://dx.doi.org/10.1080/01431161.2013.875632.

Chen, Cheng, Chen, Qiuwen, Li, Gang, He, Mengnan, Dong, Jianwei, Yan, Hanlu, Wang, Zhiyuan, Duan, Zheng, 2021. A novel multi-source data fusion method based on Bayesian inference for accurate estimation of chlorophyll-a concentration over eutrophic lakes. Environ. Model. Softw. (ISSN: 1364-8152) 141, 105057. http://dx.doi.org/10.1016/j.envsoft.2021.105057.

Chen, Cheng, Huang, Jiacong, Chen, Qiuwen, Zhang, Jianyun, Li, Zhijie, Lin, Yuqing, 2019. Assimilating multi-source data into a three-dimensional hydro-ecological dynamics model using ensemble Kalman filter. Environ. Model. Softw. (ISSN: 1364-8152) 117, 188–199.

Chiles, Jean-Paul, Delfiner, Pierre, 2012. Geostatistics: Modeling Spatial Uncertainty, second ed. In: Applied Probability and Statistics, John Wiley & Sons.

Cohn, Stephen, Sivakumaran, N., Todling, Ricardo, 1995. A fixed-lag Kalman smoother for retrospective data assimilation. Mon. Weather Rev. 122, http://dx.doi.org/10.1175/1520-0493(1994)122<2838:AFLKSF>2.0.CO;2.

Cressie, Noel, Wikle, Christopher K., 2011. Statistics for Spatio-Temporal Data. Wiley.

Crow, Wade T., 2003. Correcting land surface model predictions for the impact of temporally sparse rainfall rate measurements using an ensemble Kalman filter and surface brightness temperature observations. J. Hydrometeorol. (ISSN: 1525-755X) 4 (5), 960–973.

Doña, Carolina, Chang, Ni-Bin, Caselles, Vicente, Sánchez, Juan M., Camacho, Antonio, Delegido, Jesús, Vannah, Benjamin W., 2015. Integrated satellite data fusion and mining for monitoring lake water quality status of the albufera de valencia in Spain. J. Environ. Manag. (ISSN: 0301-4797) 151, 416–426. http://dx.doi.org/10.1016/j.jenvman.2014.12.003.

EU, 2000. The EU water framework directive - integrated river basin management for Europe. URL https://ec.europa.eu/environment/water/water-framework/index_en.html.

EU, 2008. EU marine strategy framework directive. URL https://ec.europa.eu/info/research-and-innovation/research-area/oceans-and-seas/eu-marine-strategy-framework-directive_en.

Evensen, Geir, 1994. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. J. Geophys. Res. Oceans 99 (C5), 10143–10162. http://dx.doi.org/10.1029/94JC00572, URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/94JC00572.

Evensen, Geir, 2009. Data Assimilation: The Ensemble Kalman Filter, second ed. Springer.

Fang, Shiqi, Del Giudice, Dario, Scavia, Donald, Binding, Caren E., Bridgeman, Thomas B., Chaffin, Justin D., Evans, Mary Anne, Guinness, Joseph, Johengen, Thomas H., Obenour, Daniel R., 2019. A space-time geostatistical model for probabilistic estimation of harmful algal bloom biomass and areal extent. Sci. Total Environ. (ISSN: 0048-9697) 695, 133776. http://dx.doi.org/10.1016/j.scitotenv.2019.133776.

Fasbender, D., Peeters, L., Bogaert, P., Dassargues, A., 2008. Bayesian data fusion applied to water table spatial mapping. Water Resour. Res. 44 (12), http://dx.doi.org/10.1029/2008WR006921, URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2008WR006921.

Gunia, M., 2018. EnDAS – ensemble data assimilation library. URL https://github.com/martingu11/endas.

HELCOM, 2007. Helcom Baltic sea action plan adopted on 15 november 2007 in Krakow, Poland by the HELCOM extraordinary ministerial meeting. URL https://helcom.fi/media/documents/BSAP_Final.pdf.

Intel Corporation, 2021. Intel distribution for python. URL https://www.intel.com/content/www/us/en/developer/tools/oneapi/distribution-for-python.html.

Julier, S.J., Uhlmann, J.K., Durrant-Whyte, H.F., 1995. A new approach for filtering nonlinear systems. In: Proceedings of 1995 American Control Conference - ACC'95, Vol. 3. pp. 1628–1632. http://dx.doi.org/10.1109/ACC.1995.529783.

Kallio, K., Koponen, S., Ylöstalo, P., Kervinen, M., Pyhälahti, T., Attila, J., 2015. Validation of MERIS spectral inversion processors using reflectance, IOP and water quality measurements in boreal lakes. Remote Sens. Environ. 157, 147–157.

Kalman, R.E., 1960. A new approach to linear filtering and prediction problems. J. Basic Eng. 82 (1), 35–45.

Katzfuss, Matthias, Stroud, Jonathan R., Wikle, Christopher K., 2020. Ensemble Kalman methods for high-dimensional hierarchical dynamic state-space models. J. Amer. Statist. Assoc. 115 (530), 866–885. http://dx.doi.org/10.1080/01621459.2019.1592753.

Kriebel, D., Tickner, J., Epstein, P., Lemons, J., Levins, R., Loechler, E.L., 2001. The precautionary principle in environmental science. Environ. Health Persp. 9 (109), 871–876.

Lam, Siu Kwan, Pitrou, Antoine, Seibert, Stanley, 2015. Numba: A LLVM-based python JIT compiler. In: Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC. In: LLVM '15, Association for Computing Machinery, New York, NY, USA, ISBN: 9781450340052, http://dx.doi.org/10.1145/2833157.2833162.

Ma, Pulong, Kang, Emily L., 2020. A fused Gaussian process model for very large spatial data. J. Comput. Graph. Statist. 29 (3), 479–489. http://dx.doi.org/10.1080/10618600.2019.1704293.

Melet, A., Verron, J., Brankart, J.-M., 2012. Potential outcomes of glider data assimilation in the Solomon sea: Control of the water mass properties and parameter estimation. J. Mar. Syst. (ISSN: 0924-7963) 94, 232–246. http://dx.doi.org/10.1016/j.jmarsys.2011.12.003, URL http://www.sciencedirect.com/science/article/pii/S0924796311002934.

Miyoshi, Takemasa, 2011. The Gaussian approach to adaptive covariance inflation and its implementation with the local ensemble transform Kalman filter. Mon. Weather Rev. 139 (5), 1519–1535. http://dx.doi.org/10.1175/2010MWR3570.1.

Mo, Xingguo, Chen, Jing M., Ju, Weimin, Black, T. Andrew, 2008. Optimization of ecosystem model parameters through assimilating eddy covariance flux data with an ensemble Kalman filter. Ecol. Model. (ISSN: 0304-3800) 217 (1), 157–173. http://dx.doi.org/10.1016/j.ecolmodel.2008.06.021, URL http://www.sciencedirect.com/science/article/pii/S0304380008002998.

Mouazen, Abdul M., Alhwaimel, Saad A., Kuang, Boyan, Waine, Toby, 2014. Multiple on-line soil sensors and data fusion approach for delineation of water holding capacity zones for site specific irrigation. Soil Tillage Res. (ISSN: 0167-1987) 143, 95–105. http://dx.doi.org/10.1016/j.still.2014.06.003, URL http://www.sciencedirect.com/science/article/pii/S0167198714001184.

Mourre, Baptiste, Chiggiato, Jacopo, 2014. A comparison of the performance of the 3-D super-ensemble and an ensemble Kalman filter for short-range regional ocean prediction. Tellus A: Dyn. Meteorol. Oceanogr. 66 (1), 21640. http://dx.doi.org/10.3402/tellusa.v66.21640.

Nerger, Lars, Janjic Pfander, Tijana, Schröter, Jens, Hiller, Wolfgang, 2012. A unification of ensemble square root Kalman filters. Mon. Weather Rev. 140, 2335–2345. http://dx.doi.org/10.1175/MWR-D-11-00102.1.

Nerger, Lars, Schulte, Svenja, Bunse-Gerstner, Angelika, 2014. On the influence of model nonlinearity and localization on ensemble Kalman smoothing. Q. J. R. Meteorol. Soc. 140, 2249–2259. http://dx.doi.org/10.1002/qj.2293.

Pan, Ming, Wood, Eric F., Wójcik, Rafał, McCabe, Matthew F., 2008. Estimation of regional terrestrial water cycle using multi-sensor remote sensing observations and data assimilation. Remote Sens. Environ. (ISSN: 0034-4257) 112 (4), 1282–1294. http://dx.doi.org/10.1016/j.rse.2007.02.039, URL http://www.sciencedirect.com/science/article/pii/S0034425707003367. Remote Sensing Data Assimilation Special Issue.

Park, Min Ho, Tretyakov, M.V., 2015. A block circulant embedding method for simulation of stationary Gaussian random fields on block-regular grids. Int. J. Uncertain. Quantif. (ISSN: 2152-5080) 5 (6), 527–544. http://dx.doi.org/10.1615/int.j.uncertaintyquantification.2015013781.

Pham, D.T., 2001. Stochastic methods for sequential data assimilation in strongly nonlinear systems. Mon. Weather Rev. 129, 1194–1207.

Pulliainen, Jouni, Vepsäläinen, Jenni, Kaitala, Seppo, Hallikainen, Martti, Kallio, Kari, Fleming, Vivi, Maunula, Petri, 2004. Regional water quality mapping through the assimilation of spaceborne remote sensing data to ship-based transect observations. J. Geophys. Res. Oceans (ISSN: 2156-2202) 109, C12009. http://dx.doi.org/10.1029/2003JC002167.

QGIS Development Team, 2021. QGIS Geographic Information System. QGIS Association, URL https://www.qgis.org.

Qian, Song S., Stow, Craig A., Rowland, Freya E., Liu, Qianqian, Rowe, Mark D., Anderson, Eric J., Stumpf, Richard P., Johengen, Thomas H., 2021. Chlorophyll a as an indicator of microcystin: Short-term forecasting and risk assessment in lake erie. Ecol. Indic. (ISSN: 1470-160X) 130, 108055. http://dx.doi.org/10.1016/j.ecolind.2021.108055.

Raanes, Patrick N., Bocquet, Marc, Carrassi, Alberto, 2019. Adaptive covariance inflation in the ensemble Kalman filter by Gaussian scale mixtures. Q. J. R. Meteorol. Soc. 145 (718), 53–75. http://dx.doi.org/10.1002/qj.3386.

Rauch, H.E., Tung, F., Striebel, C.T., 1965. Maximum likelihood estimates of linear dynamic systems. AIAA J. 3 (8), 1445–1450. http://dx.doi.org/10.2514/3.3166.

Revilla-Romero, Beatriz, Wanders, Niko, Burek, Peter, Salamon, Peter, de Roo, Ad, 2016. Integrating remotely sensed surface water extent into continental scale hydrology. J. Hydrol. (ISSN: 0022-1694) 543, 659–670. http://dx.doi.org/10.1016/j.jhydrol.2016.10.041.

Schabenberger, O., Gotway, C.A., 2005. Statistical Methods for Spatial Data Analysis, first ed. In: Texts in Statistical Science, Chapman and Hall/CRC.

Schroeder, T., Behnert, I., Schaale, M., Fischer, J., Doerffer, R., 2007a. Atmospheric correction algorithm for MERIS above case-2 waters. Int. J. Remote Sens. 28, 1469–1486.

Schroeder, T., Schaale, M., Fischer, J., 2007b. Retrieval of atmospheric and oceanic properties from MERIS measurements: a new case-2 water processor for BEAM. Int. J. Remote Sens. 28 (24), 5627–5632.

Seppälä, J., Ylöstalo, P., Kaitala, S., Hällfors, S., Raateoja, M., Maunula, P., 2007. Ship-of-opportunity based phycocyanin fluorescence monitoring of the filamentous cyanobacteria bloom dynamics in the Baltic sea. Estuar. Coast. Shelf Sci. 73, 489–500.

Stroud, Jonathan R., Stein, Michael L., Lesht, Barry M., Schwab, David J., Beletsky, Dmitry, 2010. An ensemble Kalman filter and smoother for satellite data assimilation. J. Amer. Statist. Assoc. 105 (491), 978–990. http://dx.doi.org/10.1198/jasa.2010.ap07636.

SYKE, 2020. Environmental information system of the finnish environment administration. URL https://www.syke.fi/en-US/Open_information.

Wang, X., Zhang, J., Babovic, V., Gin, K.Y.H., 2019. A comprehensive integrated catchment-scale monitoring and modelling approach for facilitating management of water quality. Environ. Model. Softw. (ISSN: 1364-8152) 120, 104489. http://dx.doi.org/10.1016/j.envsoft.2019.07.014, URL https://www.sciencedirect.com/science/article/pii/S1364815217307326.

Wood, Andrew T.A., Chan, Grace, 1994. Simulation of stationary Gaussian processes in [0,1]d. J. Comput. Graph. Statist. (ISSN: 10618600) 3 (4), 409–432.

Zammit-Mangion, Andrew, Cressie, Noel, Shumack, Clint, 2018. On statistical approaches to generate level 3 products from satellite remote sensing retrievals. Remote Sens. (ISSN: 2072-4292) 10 (1), http://dx.doi.org/10.3390/rs10010155.