A Python interface to the Fortran-based Parallel Data Assimilation Framework: pyPDAF v1.0.0

Yumeng Chen[1,2], Lars Nerger[3], and Amos S. Lawless[1,2]

[1]School of Mathematical, Physical and Computational Sciences, University of Reading, Reading RG6 6ET, UK
[2]National Centre for Earth Observation, University of Reading, Reading RG6 6ET, UK
[3]Alfred-Wegener-Institut, Helmholtz-Zentrum für Polar-und Meeresforschung (AWI), 27570 Bremerhaven, Germany

Correspondence: Yumeng Chen (yumeng.chen@reading.ac.uk)

Abstract. Data assimilation (DA) is an essential component of numerical weather and climate prediction. Efficient implementation of DA benefits both research and operational prediction. Currently, a variety of DA software programs are available. One of the notable DA libraries is the Parallel Data Assimilation Framework (PDAF) designed for ensemble data assimilation. The DA framework is widely used with complex high-dimensional climate models and is applied for research on atmosphere, ocean, sea ice and marine ecosystem modelling, as well as operational ocean forecasting. Meanwhile, there exists increasing need for flexible and efficient DA implementations using Python due to the increasing amount of intermediate complexity models as well as machine learning based models coded in Python. To accommodate for such needs, we introduce a Python interface to PDAF, pyPDAF. pyPDAF allows for flexible DA system development while retaining the efficient implementation of the core DA algorithms in the Fortran-based PDAF. The ideal use-case of pyPDAF is a DA system where the model integration is independent from the DA program, which reads the model forecast ensemble, produces a model analysis and updates the restart files of the model, or a DA system where the model can be used in Python. With implementations of both PDAF and pyPDAF, this study demonstrates the use of pyPDAF and PDAF for coupled data assimilation (CDA) in a coupled atmosphere-ocean model, the Modular Arbitrary-Order Ocean-Atmosphere Model (MAOOAM). This study demonstrates that pyPDAF allows for the utilisation of Python user-supplied functions with PDAF functionalities. The study also shows that pyPDAF can be used with high-dimensional systems with little slow-down per analysis step of only up to 13% for the localized ensemble Kalman filter LETKF. In addition, our CDA experiments confirm the benefit of strongly coupled data assimilation for improving both the instantaneous state and the long-term trend of the coupled dynamical system.

1 Introduction

Data assimilation (DA) is widely used in weather and climate modelling where observations are used to constrain the model prediction based on the uncertainty of both the observations and the model forecast. Due to the limited predictability and imperfect models, DA has become one of the most important techniques for the numerical weather and climate predictions. Progresses of the DA methodology development can be found in various review articles and books (e.g., Bannister, 2017; Carrassi et al., 2018; Vetra-Carvalho et al., 2018; Evensen et al., 2022).

To ameliorate the difficulties in the implementation of different DA approaches, several DA software programs and libraries have been proposed (e.g., Nerger et al., 2005; Anderson et al., 2009; Raanes et al., 2024; Trémolet and Auligne, 2020). Even though the implementation of the core DA algorithms is similar, these software programs/libraries are typically tailored to different purposes. For example, the Joint Effort for Data assimilation Integration (JEDI, Trémolet and Auligne, 2020) is a piece of self-contained software that includes a variety of functionalities that can be used for all aspects of a DA system mainly for operational purposes while DA software for methodology research such as DAPPER (Raanes et al., 2024) is designed for identical twin experiments equipped with low complexity models.

One widely used DA framework is the Parallel Data Assimilation Framework (PDAF) developed and maintained by the Alfred Wegener Institute (Nerger et al., 2005; Nerger and Hiller, 2013b). The framework is designed for efficient implementations of ensemble-based DA systems for complex weather and climate models but is also used for research on data assimilation methods with low-dimensional "toy" models. The DA implementations require user-supplied functions to provide case-specific information about the DA system including the treatment of observations and localisation. More than 100 studies have used PDAF, including atmosphere (e.g., Shao and Nerger, 2024), ocean (e.g., Losa et al., 2012; Pohlmann et al., 2023), sea ice (e.g., Williams et al., 2023; Zhao et al., 2024), land surface (e.g., Strebel et al., 2022; Kurtz et al., 2016), hydrology (e.g., Tang et al., 2024; Döll et al., 2024), and coupled systems (e.g., Nerger et al., 2020). Further use-cases of PDAF can be found in the PDAF website (PDAF - the Parallel Data Assimilation Framework, last access: 2024-02-13). Even though PDAF provides highly optimised DA algorithms, the flexible framework relies on the user-supplied functions to couple DA with model system and observations. The implementation of user-supplied functions still require additional code development, which can be time-consuming especially when the routines have to be written in Fortran, a popular programming language for weather and climate applications.

In recent years, Python is gaining popularity in weather and climate communities due to its flexibility and ease of implementation. For example, Python is adopted by some low- to intermediate-complexity models (e.g., De Cruz et al., 2016; Abernathey et al., 2022), models with a Python wrapper (e.g., McGibbon et al., 2021), and machine learning based models (e.g., Kurth et al., 2023; Lam et al., 2023; Bi et al., 2023). For the application of DA in Python, DAPPER provides a variety of DA algorithms for twin experiments using low-dimensional Python models. The Ensemble and Assimilation Tool, EAT (Bruggeman et al., 2024) was proposed to set up a 1D ocean-biogeochemical DA system, which is a wrapper to a Fortran data assimilation system based on PDAF including the 1D ocean-biogeochemical model, GOTM-FABM. There are also Python packages designed mainly for pedagogical purposes in low-dimensional systems such as openDA (Ahmed et al., 2020) and filterpy (filterpy PyPI, last access: 2024-08-29). For high-dimensional applications, there are efficient implementations of DA packages such as HIPPYlib by Villa et al. (2021) and ADAO (SALOME The Open Source Integration Platform for Numerical Simulation, last access: 2024-08-29), but HIPPYlib does not have a focus on ensemble data assimilation approaches whereas ADAO provides various ensemble DA methodologies but it has no support for the localisation used in weather and climate applications. More recently, NEDAS (Ying, 2024) was introduced for offline ensemble DA in climate applications but it currently only supports limited DA algorithms.

Targeted at applications to high-dimensional ensemble data assimilation systems, here, we introduce a Python interface to PDAF, pyPDAF. Using pyPDAF, one can implement both offline and online DA systems using Python. For offline DA systems, DA is performed utilising files written onto a disk, e.g., model restart files. If a numerical model is available in Python, pyPDAF allows for online DA system implementation where DA algorithms can be used with the Python model with in-memory data exchange that does not need I/O operations bringing about more efficiency than an offline system. Compared to user-supplied functions implemented in Fortran, the Python implementation can facilitate easier code development thanks to a variety of packages readily available in Python. In the meantime, DA algorithms provided by PDAF that are efficiently implemented in Fortran can still be utilised.

In this study, we demonstrate the use of pyPDAF in a coupled data assimilation (CDA) setup with the Modular Arbitrary-Order Ocean-Atmosphere Model (MAOOAM, De Cruz et al., 2016) where an arbitrary number of grid points can be specified without changing the model dynamics making it suitable to provide benchmarks of pyPDAF. The research on CDA is motivated by the use of coupled earth system models, especially for coupled atmosphere and ocean simulations (Eyring et al., 2016; Walters et al., 2019). Traditionally, each model component is assimilated individually and the state of each model component interacts with the others only in the coupled model forecast. This approach is called weakly coupled DA (WCDA). It is desirable to perform DA jointly for all model components simultaneously, usually denoted as strongly coupled DA (SCDA). Studies report a suite of benefits of using SCDA. For example, Smith et al. (2015) shows that the SCDA can improve dynamical balance in the analysis leading to reduced initialisation shocks. Sluka et al. (2016) reported improvements in analysis with SCDA in an intermediate complexity model. Tang et al. (2021) performed SCDA of ocean observations into the coupled atmosphere-ocean model AWI-CM and found positive effects in particular in the polar regions. Further studies can be found in a suite of review articles on CDA (Penny and Hamill, 2017; Zhang et al., 2020; de Rosnay et al., 2022; Kalnay et al., 2023).

Here, we will first introduce ensemble-based data assimilation, the principal objective of PDAF, in Sect. 2. Section 3 will describe the design and implementation of PDAF and pyPDAF. In Sect. 4, the experimental and model setup will be described. Section 5 will report the performance of PDAF and pyPDAF in CDA setup. We will conclude in Sect. 6.

2 Ensemble-based data assimilation

Although PDAF supports a few deterministic DA methods, it focuses on ensemble-based DA methods. Ensemble-based DA is a class of DA approaches that approximate the statistics of the model state and its uncertainty using an ensemble of model realisations motivated by DA approaches based on Bayes theorem where the prior, typically a model forecast, and posterior (analysis) distributions can be approximated by a Monte Carlo approach. The ensemble model forecast allows for an embarrassingly parallel implementation which means that, with sufficient computational resources, the wall clock computational time of the forecast does not increase with the ensemble size.

Under the Gaussian assumption of the forecast and analysis distributions, one of the most notable ensemble-based DA methods is the ensemble Kalman filter (EnKF, Evensen, 1994). The EnKF approximates the forecast and analysis error distribution by an ensemble. The method was proven to be successful in many applications (e.g., Houtekamer et al., 2005; Feng et al., 2009;

Hamill et al., 2011; Sakov et al., 2012). To further improve the efficiency and reliability of the EnKF, multiple variants of the EnKF were proposed, such as singular evolutive intepolated Kalman filter (SEIK, Pham, 2001), ensemble transform Kalman filter (ETKF, Bishop et al., 2001), error space transform Kalman filter (ESTKF, Nerger et al., 2012), and the deterministic ensemble Kalman filter (Sakov and Oke, 2008). In practice computational resources limit the feasible ensemble size, which is typically of an order of 10 to 100, in the high-dimensional realistic DA applications in the Earth system due to the cost of model forecasts. The ensemble-based DA approaches typically suffer from sampling errors from limited ensemble size. To counter these deficiencies, covariance matrix inflation and localisation are commonly used (e.g., Pham et al., 1998; Hamill et al., 2001; Hunt et al., 2007). In particular, the domain localisation is tailored for efficient parallel implementations that are commonly used in high-dimensional DA systems.

Ensemble-based DA can also treat fully non-linear non-Gaussian problems. The most notable example is particle filters (see, van Leeuwen et al., 2019). They can be used to solve fully non-linear problems without assumptions on the prior and posterior distribution. However, for high-dimensional geoscience applications, the classical particle filters suffer from the "curse of dimensionality" where the required ensemble size grows exponentially with the dimension of the state vector making the approach computationally infeasible. Recent developments of the particle filters significantly improve the stability and reduce the required ensemble size of the approach making it a potential choice for low-to-medium complexity models, such as implicit equal-weights particle filters (Zhu et al., 2016) and the particle flow filter (Hu and van Leeuwen, 2021). An overview of other developments of particle filters can be found in van Leeuwen et al. (2019).

The ensemble-based DA approaches are adopted by many operational centres where traditionally variational methods are used (e.g., Clayton et al., 2013; Caron et al., 2015; Bonavita et al., 2016; Hersbach et al., 2020). In variational methods, ensemble approaches are used to achieve flow-dependent background covariance matrix, and/or to avoid explicit computation of the adjoint model in the minimisation process by using an ensemble approximation. These goals can be realised using various different methodologies and a detailed review of these methods can be found in Bannister (2017).

3 PDAF and PyPDAF

PDAF is designed for research and operational DA systems. As a Python interface to PDAF, pyPDAF inherits the DA algorithms implemented in PDAF and the same implementation approach to build a DA system.

3.1 Parallel Data Assimilation Framework (PDAF)

PDAF is a Fortran-based DA framework providing fully optimised, parallelised ensemble-based DA algorithms. The framework provides a software library and defines a suite of workflows based on different DA algorithms provided by PDAF including various ensemble Kalman filters/smoothers, ensemble-based 3DVar (Bannister, 2017), particle filters (van Leeuwen et al., 2019) and other non-linear filters (Tödter and Ahrens, 2015; Nerger, 2022). To deal with sampling errors in the ensemble-based DA, the framework also provides options for adaptive inflation and localisation schemes.

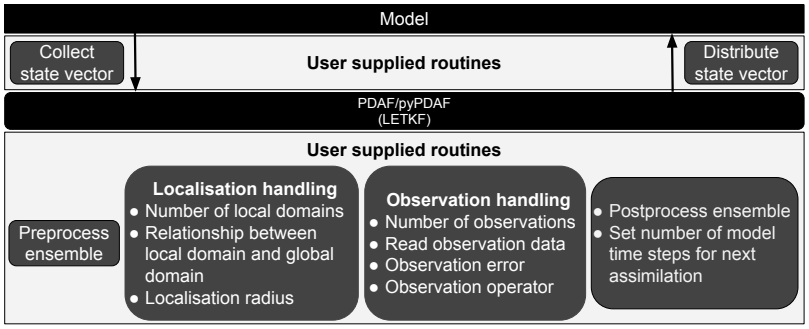

Figure 1. A schematic diagram of an online LETKF DA system using (py)PDAF. In the case of an offline DA system, the model can be its restart files.

As a framework for ensemble DA, it comes with the functionality to generate the initial ensemble. One possibility is to use the second-order exact sampling (Pham, 2001) where the ensemble is generated based on the model trajectory of the modelled truth. The assumption is that the uncertainty of the model initial condition lies in the phase space of the model trajectory. The space is represented by the singular values and its corresponding vectors using an empirical orthogonal function (EOF) decomposition.

To ensure that PDAF can be flexibly adapted to any models and observations, it requires users to provide case-specific information. This includes the information on the state vector, observations and localisation. The framework obtains this information via *user-supplied functions* which are external callback subroutines. Figure 1 shows a schematic diagram of an online DA system where the LETKF is used. Here, the user-supplied functions connect PDAF with models. Called within the PDAF routines, these user-supplied functions collect state vectors from model forecasts and distribute the analysis back to the model for the following forecast phase. During the analysis step, user-supplied functions also pre- and post-process the ensemble, handle localisations and observations, and provide the number of model time steps for the next forecast phase to PDAF. As the user-supplied functions depend on the chosen DA algorithm, other algorithms may require different functions. For example, the 3DVar requires routines for the adjoint observation operator and control vector transformation. To ameliorate the difficulty in the observation handling, PDAF provides a scheme called observation module infrastructure (OMI). The OMI routines handle the processing of observation vectors and error covariance matrix used by DA algorithms, and provide support for the complex distance computation used by localisation. In the current version of PDAF V2.3, it also supports spatial interpolations on structured and unstructured grids, direct observation operator, and a diagonal or non-diagonal observation error covariance matrix. One can also implement PDAF without OMI, but additional functions would be required.

In an online DA system, the collection and distribution of state vector is an in-memory data exchange handled by PDAF efficiently. It is possible to implement an offline DA system with PDAF where the model in Fig. 1 would be replaced by model restart files while the user-supplied collection and distribution routines manage the I/O operations of these restart files. Offline DA implementation is a crucially supported feature of PDAF and a potentially important use-case for pyPDAF, but we will

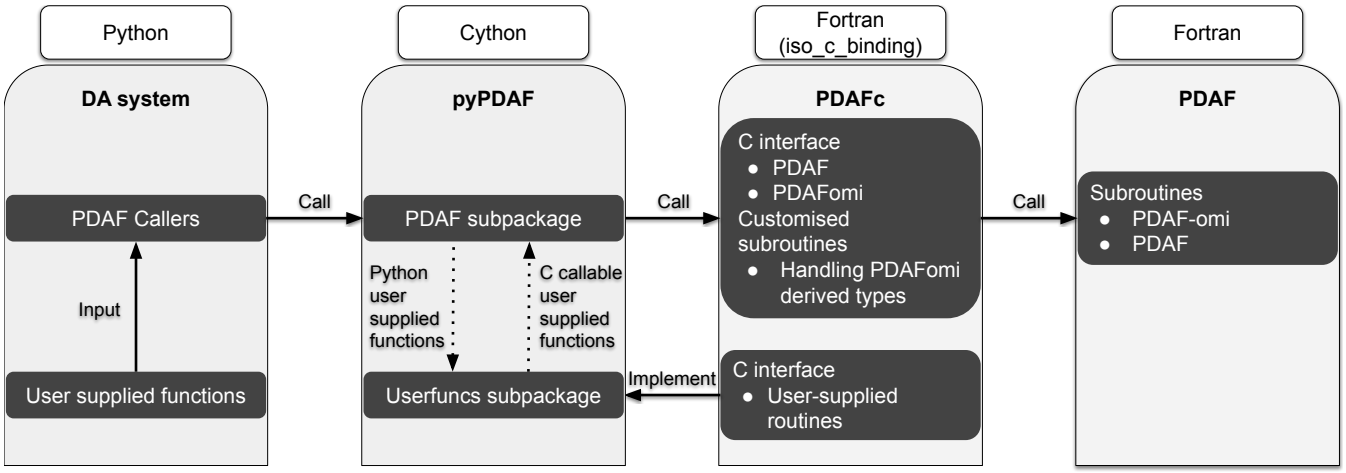

Figure 2. An illustration of the design of the pyPDAF interface to the Fortran-based framework PDAF. Here, only the Python component is exposed to pyPDAF users, and the Cython and Fortran implementations are internal implementations of pyPDAF.

not discuss it in detail for the sake of brevity. We will provide details of the use of user-supplied functions in the context of pyPDAF in Sect. 3.3.

3.2 pyPDAF

Implementation of user-supplied functions can be laborious in Fortran and typical code development in Python can be less time consuming. Thanks to the integrated package management, code development in Python can rely on well optimised packages without the need for compilation. For these reasons, a variety of numerical models are implemented in Python (e.g., De Cruz et al., 2016; Abernathey et al., 2022; McGibbon et al., 2021; Bi et al., 2023). Hence, a Python interface to PDAF allows the design of an online DA system with such Python-based models. These range from low-dimensional toy dynamical systems to high-dimensional weather and climate systems. Compared to a Fortran-coded DA system, a Python DA system can be implemented efficiently and allows for easier modifications such that users can focus on scientific problems.

The pyPDAF package can also be applied for offline DA systems, i.e. coupling the model and data assimilation program through restart files. Here pyPDAF can be used without the restriction of the programming language of the numerical model. When computation-intensive user-supplied functions are well optimised (e.g., using just-in-time (JIT) compilation), this could also be used for complex models. Thus, depending on the requirements of the users, an offline DA system can be used to prototype a Fortran DA system as well. The application of pyPDAF in high-dimensional models can also be shown by its support of the parallel features of PDAF, which use the Message Passing Interface (MPI, Message Passing Interface Forum, 2023). For this, a pyPDAF DA system relies on the "mpi4py" package for MPI support. The pyPDAF system can also support shared memory parallelisation in PDAF when built with OpenMP.

As the reference implementation of Python is based on the C programming language (The Python Language Reference, last access: 2024-02-13), the design of pyPDAF is based on the interoperability between the programming languages of C and Fortran using the *iso_c_binding* module of Fortran. As shown in Fig. 2, the C interface of PDAF, *PDAFc*, is developed in pyPDAF, which includes essential PDAF interfaces and interfaces for user-supplied functions. Hence, PDAFc could be used independently from pyPDAF as a C interface to the PDAF package. The core of the pyPDAF implementation uses the C-extension for Python (Cython). Here Python datatypes are converted into C pointers to allow for information exchange between PDAF and pyPDAF. pyPDAF implements C callable functions which can call user-supplied functions in Python such that PDAF can utilise the user-supplied Python functions.

pyPDAF is designed so that a DA system can be coded purely in Python including the user-supplied functions and function calls to algorithms implemented in PDAF. The interface to PDAF is provided through functions implemented using Cython, which provides the interface for calls from Python. Thus, the pyPDAF package itself is a mixed program of C, Fortran and Python. Moreover, as DA algorithms require high-dimensional matrix multiplications, PDAF relies on the numerical libraries LAPACK (linear algebra package) and BLAS (basic linear algebra subprograms). These libraries lead to a complex compilation process especially when users could use different operating systems. To fully utilise the cross-platform support of Python environment, pyPDAF is distributed via the package manager *conda* to provide an out-of-box user experience with pyPDAF where users can use pyPDAF without the need for compiling the package from the source code. Detailed installation instructions can be found at: https://yumengch.github.io/pyPDAF/install.html.

pyPDAF allows for the use of efficient DA algorithms in PDAF. However, a DA system purely based on pyPDAF could still be less efficient than a DA system purely based on PDAF coded in Fortran. The loss of efficiency is partly due to the operations in user-supplied Python functions and the overhead from the conversion of data types between Fortran and Python. We will evaluate the implications of these loss of efficiency in Sect. 5.2.

3.3 Construction of data assimilation systems using pyPDAF

To illustrate the application of pyPDAF to an existing numerical model, as an example, we present key components of an LETKF DA system. This example follows the schematic diagram in Fig. 1. Here, we assume that the number of processors is equal to the ensemble size. In this setup, each ensemble member of the model forecast runs on one processor, and the analysis is performed serially on a single processor. We further assume that observations are co-located on the model grid but are of lower resolution, and they have a diagonal error covariance matrix.

Compared to Fortran implementations, a Python DA system can better utilise the object-oriented features. Here, we assume the existence of a generic *model* object that contains model information. In this sytem, the pyPDAF functionalities should be

initialised by

$$param\_int, param\_real, flag = pyPDAF.PDAF.init(filtertype, subtype, stepnull, \\ param\_int, param\_real, \\ COMM\_model, COMM\_filter, COMM\_couple, \\ task\_id, n\_modeltasks, filterpe, init\_ens\_pdaf).$$

The information on the type of filters (*filtertype* and *subtype*) is given to PDAF by this function. It also takes parameters of these filters. Here, the size of the state vector (*dim_p*) and the ensemble size (*dim_ens*) are specified in the *param_int* array, and the inflation factor is specified in *param_real* array. These parameters allow PDAF to allocate arrays such as the ensemble mean (*state_p*) and the ensemble matrix (*ens_p*) used by the DA. The MPI communicators of model, the filter and the coupling between model and filter are also specified here by *COMM_model*, *COMM_filter*, *COMM_couple* respectively. The initialisation function also obtain other parallelisation information from the function call including the index of the parallel model tasks by *task_id*, the total number of parallel model tasks by *n_modeltasks*, a boolean variable that determine if the filter is performed on current process by *filterpe*. Detailed explanations of the parallelisation strategy used by PDAF can be found in Nerger and Hiller (2013a). Also, the initialisation function takes the initial time step, *stepnull*, as a step counter in PDAF. In the initialisation, a user-supplied function of *state_p, uinv, ens_p, flag = init_ens_pdaf(filtertype, dim_p, dim_ens, state_p, uinv, ens_p, flag)* is used to initialise PDAF ensemble, *ens_p*. In the user-supplied function, the input arguments are given by the PDAF, and the returned arguments are received by PDAF to perform DA. Here, *uinv* is a variable used for the second-order exact sampling. The ensemble generation method can be used with *pyPDAF.PDAF.eofcovar* and *pyPDAF.PDAF.SampleEns* when starting from a deterministic run. If OMI is used, *pyPDAF.PDAF.omi_init(n_obs)* is used to initialise the *n_obs* types of observations.

In each model integration step, the analysis step is executed by

$$status = pyPDAF.PDAF.omi\_assimilate\_local(collect\_state, distribute\_state, init\_dim\_obs, \\ obs\_op, prepostprocess, init\_n\_domains, \\ init\_dim\_l, init\_dim\_obs\_l, \\ g2l\_state, l2g\_state, next\_observation)$$

where *status* is a flag for the error code of the DA step, and the arguments of *pyPDAF.PDAF.omi_assimilation_local* are user-supplied functions, which will be discussed in detail. In the analysis step, each user-supplied function will next be executed by PDAF to collect necessary information, or perform case-specific operations for the DA. A flow chart is given in Fig 3.

As shown in Fig. 1, the model and PDAF exchanges information by user-supplied functions. The user-supplied function *state_p = collect_state(dim_p, state_p)* is executed by PDAF for each ensemble member to fill model forecast fields into a one-dimensional array, *state_p*. Similarly, *state_p = distribute_state(dim_p, state_p)* distributes analysis (*state_p*) to model fields for the initialisation of the next forecast cycle. These user-supplied functions allow users to adapt a DA system with different models.

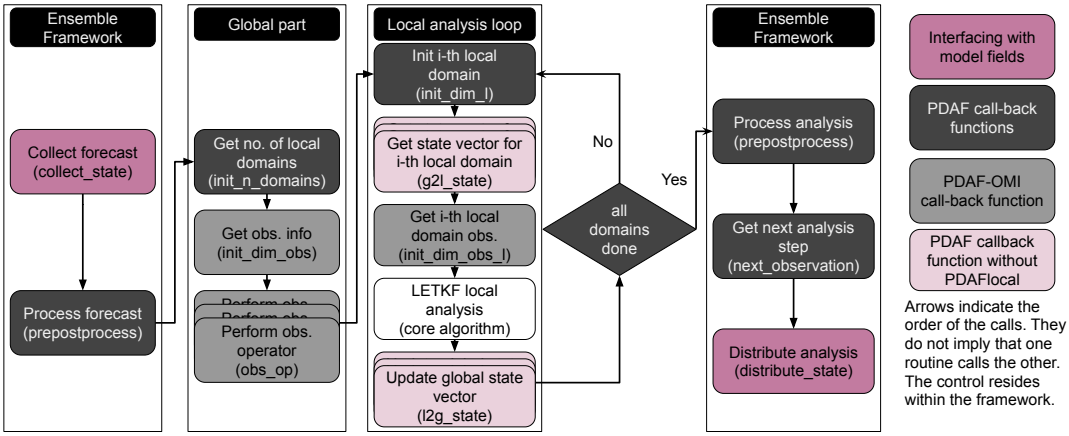

Figure 3. A flowchart of the sequence of LETKF operations in PDAF. These operations include user-supplied functions and core LETKF algorithm. The arrows indicate the order in which the user-supplied functions are executed. They do not imply that one routine calls the other. The observation operators and the global and local domain update are represented by multiple boxes as they are performed by each ensemble member.

To handle different observations, with the OMI functionality, only three user-supplied functions need to be implemented. One is *dim_obs = init_dim_obs(step, dim_obs_p)*. The primary purpose of the function is to obtain the dimension of observation vector, *dim_obs*, with an initial dimension given by *dim_obs_p* at the current time step, *step*, as implied by its name. In this function, one has to provide further observation information to OMI. The OMI obtains the information in two approaches. One approach is by calling the function: *dim_obs = pyPDAF.PDAF.omi_gather_obs(i_obs, obs_p, ivar_obs_p, ocoord_p, cradius)*. The function returns the total dimension of the observation vector (dim_obs) of *i_obs*-th observation type which is returned by the user-supplied function *init_dim_obs*. As function arguments, *pyPDAF.PDAF.omi_gather_obs* provides PDAF with the observation vector (*occord_p*), inverse of the observation variance (*ivar_obs_p*), the observation coordinates (*ocoord_p*), and a localisation radius for the current observation type (*cradius*). The other approach sets attributes of the derived data type, *obs_f*, in PDAF. In *obs_f*, the attributes include the switch of the assimilation of the observation type, the index of the observation in the state vector, *id_obs_p*, the domain size and the options for distance computation in localisation. While these attributes can be set by direct initialisation in Fortran, in pyPDAF, these attributes can be set by setter functions, e.g., *id_obs_p* can be set using the pyPDAF function *pyPDAF.PDAF.omi_set_id_obs_p(i_obs, id_obs_p)*.

The observation operator is implemented by the user-supplied function *m_state_p = obs_op(step, dim_p, dim_obs_p, state_p, m_state_p)*. It takes a state vector (*state_p*) as input and returns a vector in observation space (*m_state_p*). In our example, it can be handled directly by the OMI function *m_state_p = pyPDAF.PDAF.omi_obs_op_gridpoint(i_obs, state_p, m_state_p)*. Note that other observation operators are also available with pyPDAF but not discussed here. The last user-supplied function related to observations is *dim_obs_l = init_dim_obs_l(domain_p, step, dim_obs, dim_obs_l)* which tells PDAF the number of observations being assimilated in the current local domain (*dim_obs_l*). This function can be simplified by the OMI function *dim_obs_l = pyPDAF.PDAF.omi_init_dim_obs_l_iso(i_obs, coords_l, locweight, cradius, sradius, dim_obs_l)* which automatically han-

dles observation vectors and its error variances used in the local domain given the coordinate of local domain (*coords_l*), the type of localisation weight (*locweight*), and the localisation radius (*cradius*) as well as the support radius of localisation function (*sradius*).

The domain localisation requires four additional user-supplied functions. The number of local domains (*n_domains_p*) is provided by *n_domains_p = init_n_domains(step, n_domains_p)*, the dimension of *domain_p*-th local domain, *dim_l*, is provided by *dim_l= init_dim_l(step, domain_p, dim_l)*. The conversion of the full global state vector to a state vector on local domain and vice versa is controlled by*state_l = g2l_state(step, domain_p, dim_p, state_p, dim_l, state_l)* and *state_p = l2g_state(step, domain_p, dim_l, state_l, dim_p, state_p)*. The user-supplied function *g2l_state* and *l2g_state* are not used in 'PDAFlocal' modules as will be discussed in Sect. 5.2.

The pyPDAF analysis step requires two additional user-supplied functions. The *state_p, uinv, ens_p = prepostprocess(step, dim_p, dim_ens, dim_ens_p, dim_obs_p, state_p, uinv, ens_p, flag)* function is called by PDAF to preprocess the forecast ensemble (*ens_p*) before the LETKF and post-process the analysis ensemble (*ens_p*) after the LETKF assimilated the observations. The user-supplied function, *nsteps, doexit, time = next_observation(step, nsteps, doexit, time)*, tells PDAF the number of time steps between two DA executions, *nsteps*. Given the current time step and other uninitialised input arguments, PDAF also obtains the information of the current model time, *time* and a flag for the completion of all DA cycles *doexit* in *next_observation*. To control the memory allocation in the DA cycle, the DA system can be finalised by function *pyPDAF.PDAF.deallocate()*.

PDAF can handle much more complex cases including non-isotropic localisation, or non-diagonal observation error covariance matrices. Besides LETKF, other filters might require different user-supplied functions as they utilise different case-specific information. The code that exists can support a wide range of filters without changes.

4 Model and DA setup

To demonstrate the application of pyPDAF and to evaluate its performance in a coupled DA setup, MAOOAM (De Cruz et al., 2016) version 1.4 is coupled with PDAF and pyPDAF. The original MAOOAM model is implemented in Fortran that is coupled directly with PDAF, and a wrapper for Python is developed in this study such that MAOOAM can be coupled with pyPDAF. This means that two online DA systems using Fortran and Python respectively are developed to allow for a comparison between the PDAF and pyPDAF implementation. In these DA systems, a suite of twin experiments is carried out using the ensemble transform Kalman filter (ETKF, Bishop et al., 2001) and its domain localisation variant, LETKF.

4.1 Coupled model MAOOAM

The MAOOAM solves a reduced-order non-dimensionalised quasi-geostrophic (QG) equation (De Cruz et al., 2016). Using the beta-plane approximation, the model has a two-layer QG atmosphere component and one-layer QG shallow-water ocean component with both thermal and mechanical coupling. For the atmosphere, the model domain is zonally periodic and has a no-flux boundary condition meridionally. For the ocean, no-flux boundary conditions are applied in both directions. This setup represents a channel in the atmosphere and a basin in the ocean. The model variables for the two-layer atmosphere are averaged

into one layer. Accordingly, MAOOAM consists of four model variables: the atmospheric streamfunction, ψ_a, the atmospheric temperature, T_a, the ocean streamfunction, ψ_o, and the ocean temperature, T_o. The model variables are solved in a spectral space. The spectral basis functions are orthonormal eigenfunctions of the Lapace operator subject to the boundary condition, and the number of spectral modes is characterised by harmonic wave numbers P, H, M (Cehelsky and Tung, 1987).

We integrate MAOOAM with (py)PDAF. As shown in Fig. 1, the key ingredient of coupling MAOOAM with (py)PDAF is the collection and distribution of state vector. In common setups of ocean and atmospheric DA, the observations are available in the physical space. Hence, in the user-supplied function that collects the state vector for pyPDAF (see Fig. 1), spectral modes of the model are transformed from the spectral space to physical space using the transformation equation,

$$f(x,y,t) = \sum_{i=1}^{K} c_i(t) F_i(x,y), \tag{1}$$

where $f(x,y,t)$ is any model variable in the physical space, K is the number of modes, $c_i(t)$ is the spectral coefficient of the model variable, $F_i(x,y)$ is the spectral basis function of mode i outlined in De Cruz et al. (2016). In the user-supplied function that distributes the state vector for pyPDAF (see Fig. 1), the analysis has to be transformed back to the spectral space to initialise the following model forecast. The inverse transformation from the physical space to the spectral space can be obtained by

$$c_i(t) = \frac{n}{2\pi^2} \int_0^{\pi} \int_0^{\frac{2\pi}{n}} f(x,y,t) F_i(x,y) dx dy. \tag{2}$$

Here, each basis function corresponds to a spectral coefficient of the model variable. The basis functions are evaluated on an equidistant model grid. The spectral coefficients are obtained via the Romberg numerical integration. To ensure the accuracy of the numerical integration, the number of grid points is $2^k + 1$ with $k \in \mathcal{Z}^+$.

Our model configuration adopts the strongly coupled ocean and atmosphere configuration (36st) of Tondeur et al. (2020) using a time step of 0.1 time units corresponding to around 16 minutes. Using the notation of $H^{max}x - P^{max}y$ of De Cruz et al. (2016) with the superscript max the maximum number of harmonic wave numbers, the configuration chooses $2x - 4y$ modes for the ocean component and $2x - 2y$ modes for the atmosphere component. This leads to a total of 36 spectral coefficients with 10 modes for ψ_a and T_a respectively and 8 modes for ψ_o and T_o respectively. The model runs on a rectangular domain with a reference coordinate system of $(x \times y) \in [0, \frac{2\pi}{n}] \times [0, \pi]$, where $n = 1.5$ is the aspect ratio between the x and y dimensions.

In contrast to Tondeur et al. (2020) who assimilate in the spectral space, we assimilate in the physical space in which the observations are usually available. A sensitivity experiment was performed to study the transformation error. The experiment shows that when the number of grid points reaches $(2^7 + 1 \times 2^7 + 1) = (129 \times 129)$, the transformation error becomes negligible and the physical grid points resolve the features in the spectral space. In practice, due to the chaotic nature of the model and long simulation time, the error from the transformation can accumulate which subsequently leads to model errors. The transformation between the spectral and physical space allows us to investigate the computational cost of the DA in pyPDAF and PDAF with the same model dynamics. As the ensemble size is determined by the dimension of unstable subspace of the dynamical system, a fixed ensemble size can be used (Tondeur et al., 2020). Therefore, for benchmarking computational cost, we conduct a suite of SCDA experiments with $2^k + 1 \times 2^k + 1$ number of grid points where $7 \leq k \leq 11$. This gives us state

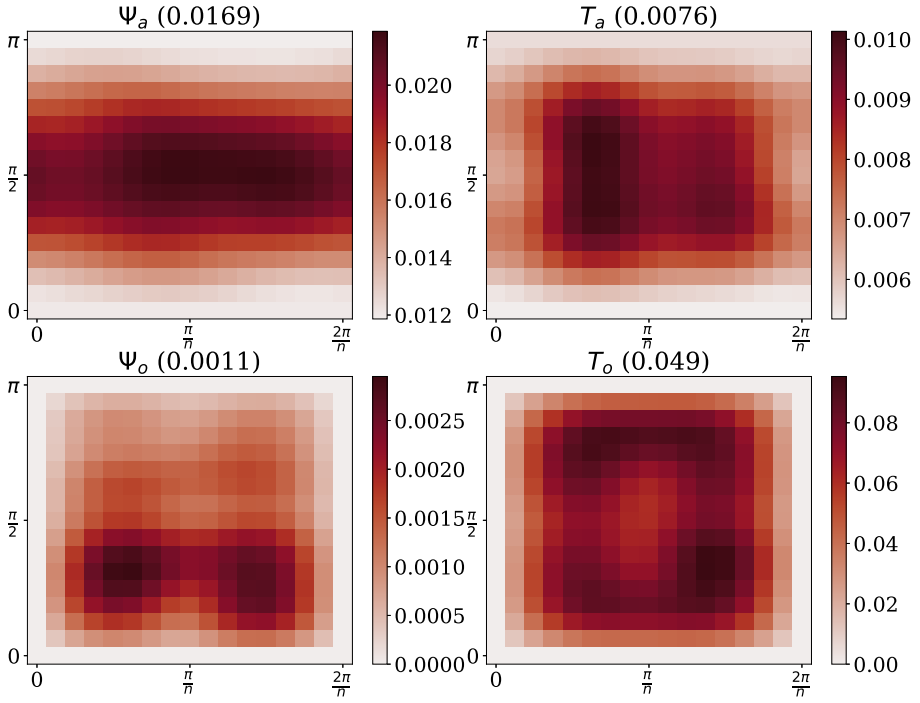

Figure 4. The observation error standard deviation fields used for generating the synthetic observations. The spatial mean of the error standard deviation is shown in the bracket.

vectors with dimension ranging from a magnitude of 10^4 to 10^7. The size of a state vector with around 10^7 elements is closer to operational setups. We also implement SCDA experiments using LETKF on a grid number of 257×257 with observations on every 4 and 8 grid points to investigate the efficiency of the domain localisation in pyPDAF.

4.2 Experiment design

In a twin experiment, a long model run is considered truth. The model state is simulated with an initial condition sampled in the spectral space which follows a Gaussian distribution, $\mathcal{N}(0, 0.01)$. The DA experiments are started after 10^5 time steps corresponding to around 277 years of model integration to ensure the dynamical consistency of the model state.

The observations are generated from the truth of the model state based on pre-defined error statistics of the observations. Except for the LETKF experiments, both atmosphere and ocean observations are sampled every 8 model grid points for each model grid setup. In all cases, the observation error standard deviations are set to 50% and 70% of the temporal standard deviation of the true model trajectory for the atmosphere and ocean respectively. The resulting standard deviation of the atmosphere observations is on a similar magnitude with the ensemble spread of the free run (cf. Fig. 5) while the magnitude of the observation error in the ocean is typically larger than in the atmosphere in real observing networks. As an example, the obtained standard deviation fields on a grid with 17×17 grid points are shown in Fig. 4. With our chosen model configuration,

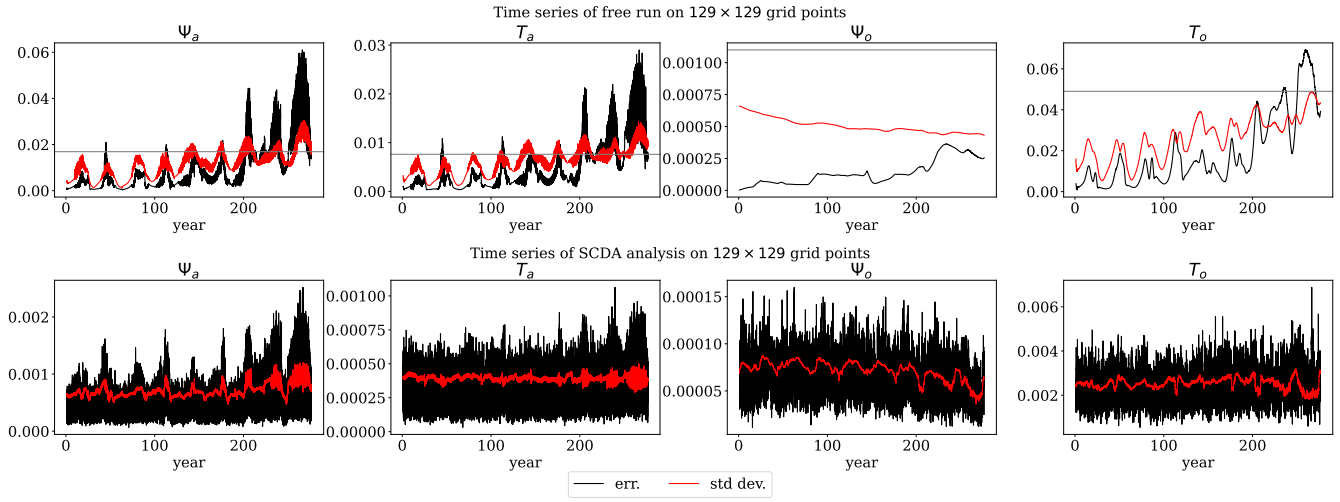

Figure 5. Ensemble spread and RMSE of the (top) free run and (bottom) SCDA analysis on a 129×129 grid. Shown are the time series of the spatial mean of ensemble spread (red) and the RMSE of the analysis (black). The atmosphere shows fast variability and oscillatory RMSE of the ensemble mean while the RMSE of the ocean ensemble mean is smooth.

the highest observation error is in the ocean temperature while the ocean streamfunction shows the least uncertainty due to its slow variability. The atmospheric processes in MAOOAM show variability on shorter timescales than the ocean. Hence, the ocean observations are assimilated around every 7 days (630 time steps) while the atmosphere observations are assimilated around every day (90 time steps).

As shown by Tondeur et al. (2020), DA in the model configuration using 36 spectral coefficients can achieve sufficient accuracy with 15 ensemble members. In this study, 16 members are used and each ensemble member runs serially with a single process. Without tuning, a forgetting factor of 0.8 is applied to maintain the ensemble spread and ensure a stable DA process.

Using the second-order exact sampling provided by PDAF (see Sect. 3.1), the ensemble is generated from a model trajectory by sampling the modelled truth every 10 days over 100 years after around 1000 years from the beginning of the simulation. This leads to 36 non-zero singular values equaling to the number of spectral modes in the model. The perturbation from the second-order exact sampling could violate the dynamical consistency of the model, so that the ensemble would need to be spun up to reach dynamical consistency. To reduce the spin up time, the initial perturbation is scaled down by a factor of 0.2, 0.15, 0.4 for Ψ_a, T_a and T_o respectively. Because the ocean streamfunction has very low variability, its perturbation is unchanged.

The DA experiments are started after 15 days from the beginning of the ensemble generation. In this setup, the forecast error is solely a result of inaccuracy of initial conditions. As shown in Fig. 5, the ensemble spread generally captures the trend and is in a similar magnitude of the model forecast error. This suggests that the forecast uncertainty from the free run ensemble initialised by the second-order exact sampling is able to reflect the forecast errors even though the spread is lower than the RMSE after 200 years.

In the free run (upper panel of Fig. 5), the ocean temperature shows the highest uncertainty of all model variables. The ocean streamfunction shows a very slow error growth rate. This is also shown by the error and ensemble uncertainty which are two magnitudes smaller than those of other model variables. Sensitivity tests (not shown) suggest that an increased error of the ocean streamfunction has a significant impact on the model dynamics consistent with the theoretical discussion given in Tondeur et al. (2020). The error of the atmosphere components shows a wave-like behaviour in time. Tondeur et al. (2020) describe the periods associated with fast dynamics with high and oscillatory errors as active regimes and the periods associated with slow dynamics with low and stable errors as passive regimes.

5 Results

In this section, we evaluate the DA skill of the MAOOAM-(py)PDAF online DA system using the ETKF. For the sake of efficiency, the skill of DA is assessed on a domain with 129×129 grid points. To evaluate the computational efficiency of pyPDAF and PDAF and the potential practical applications of pyPDAF, we compare the wallclock time in the SCDA system. The online DA systems using PDAF and pyPDAF produce quantitatively the same results in all experiments up to machine precision.

5.1 Effect of coupled data assimilation

In WCDA, the coupling only occurs during the model forecast. This means that the observations only influence their own model component in the analysis step. In this setup, each model component has its own DA system with only two model variables, the streamfunction and temperature, on the same model grid. This implies two separate DA systems. In an online DA setup in PDAF, two separate state vectors have to be defined in each analysis step which is not straightforward with PDAF due to its assumption that each analysis step has only one state vector. In the case of AWI-CM in Tang et al. (2021), two separate state vectors were obtained by using a parallelisation, but here the two model components of MAOOAM are run using the same processor. In our implementation we obtain WCDA by resetting the time step counter in PDAF in our implementation such that even if the assimilation of two state vectors are done by using PDAF twice, PDAF only counts it as one analysis time step. An alternative approach could be to use the LETKF method and define the local state vector as either the atmosphere or ocean variables.

Figure 6 shows that the time averaged RMSE of WCDA is smaller than that of the unconstrained free run. Thus, the error growth is successfully controlled. This also demonstrates that the ETKF leads to a converged analysis even though our observations are less accurate than the forecast at the start of the DA period. The results also show that sparse observations can successfully control errors in regions without observations. This is due to the fact that the model fields are rather smooth.

Compared to the WCDA, atmosphere observations influence the ocean part of the state vector and vice versa in the SCDA. This means that the coupling occurs for both the analysis step and model forecast. In this case, the DA system only has one unified state vector that contains the streamfunction and temperature of both model components. The implementation of an online SCDA system aligns with the design of PDAF and does not require special treatment.

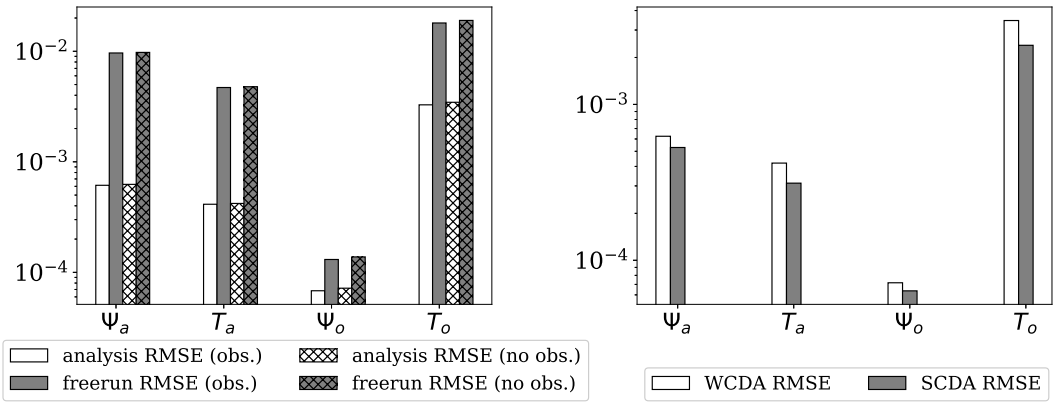

Figure 6. Left: The time-averaged RMSE of the analysis using WCDA and free run where the RMSE of the observed (non-hatched bars), denoted by "obs." in the legend, and unobserved gridpoints (hatched bars), denoted by "no obs.", are compared separately. Right: comparison of RMSEs for weakly and strongly coupled DA for all grid points. The y-axis is plotted in the log-scale.

As expected, the SCDA yields lower analysis errors than the free run as shown in Fig. 5, and the errors are also lower than the WCDA as shown in the right panel of Fig. 6. The improved analysis in the SCDA in each model component is a result of assimilating observations from the other model component. The effective use of these additional observations relies on the error cross-covariance matrix between model components estimated by the forecast ensemble. The improvements suggest a reliable error cross-covariance matrix in the coupled DA system.

To further show the performance of pyPDAF in a SCDA setup, we carry out experiments in which only one model component is observed. In the SCDA, the analysis increment of a model component without observations relies on the error cross-covariance matrix with the model components that have observations. In this experiment, inflation is only applied to the observed model component to avoid excessive analysis increment to the unobserved model components. The partial inflation is achieved in the post-processing routines as PDAF applies inflation uniformly to the entire state vector by default.

Figure 7 shows the time-averaged RMSE of fields that are smoothed in time by a moving average as a function of the averaging time-window. The RMSEs of the instantaneous model fields are represented by zero moving average window length. Assimilating observations from the other model component with SCDA can improve the analysis of the unobserved model component. The assimilation not only improves the instantaneous model fields but also the long-term trend of the atmosphere and ocean climate even though the error dynamics of atmosphere and ocean shows strong time-scale differences in Fig. 5. This means that the ocean dynamics benefit from atmosphere observations even if the transient atmosphere processes are smoothed by the moving average. Notably, the RMSE of the ocean streamfunction when only atmosphere observations are assimilated does not decrease monotonically with the moving average window length. This could be explained by the fact that the time averaged ocean streamfunction shows periodic features in time and an moving average of ~ 60 years leads to a time series of nearly constant streamfunction. This improves the skill of the DA. However, this feature is not captured by the analysis that assimilates ocean observations perhaps due to the large observation uncertainties.

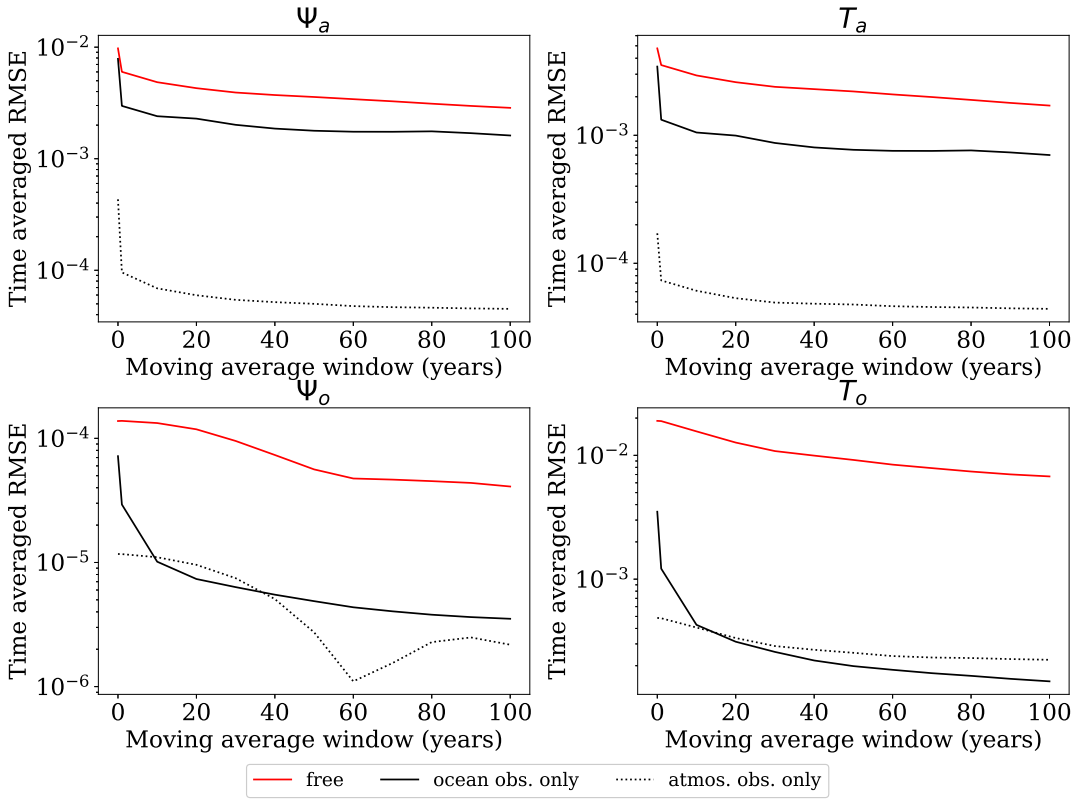

Figure 7. Time averaged RMSE when only one model component is observed. The y-axis is in log-scale.

5.2 Computational performance of PDAF and pyPDAF

One motivation of developing a Python interface to PDAF is that the efficient DA algorithms in PDAF can be used by pyPDAF while the user-supplied functions can be developed with the ease of Python. However, the user-supplied functions provided by Python are expected to be slower than a pure Fortran implementation. The slow-down is both a result of lack of compilation in Python and the type cast between Fortran arrays and Python objects. Here we present a comparison of the wall clock time of both PDAF and pyPDAF experiments with standard SCDA broken down to the level of subroutines. Each experiment runs 100 analysis steps and each experiment is repeated 10 times. The computation runs on the computing facility of University of Reading on a node with two AMD EPYC 7513 32-Core processors which have a 2.6GHz frequency. With 16 ensemble members, each member uses a single processor for model forecast and the DA is performed serially on a single processor.

As shown in Fig. 8, the PDAF-internal procedures (labeled 'internal'), which are the core DA algorithm, take nearly the same amount of time per analysis step for PDAF and pyPDAF regardless of the number of grid points. As expected, the user-supplied functions take more computational time in the DA system based on pyPDAF than PDAF. In this study, the pre- and post-processing of the state vector (labeled 'pre-post') calculates the square root of the spatial mean of ensemble variance.

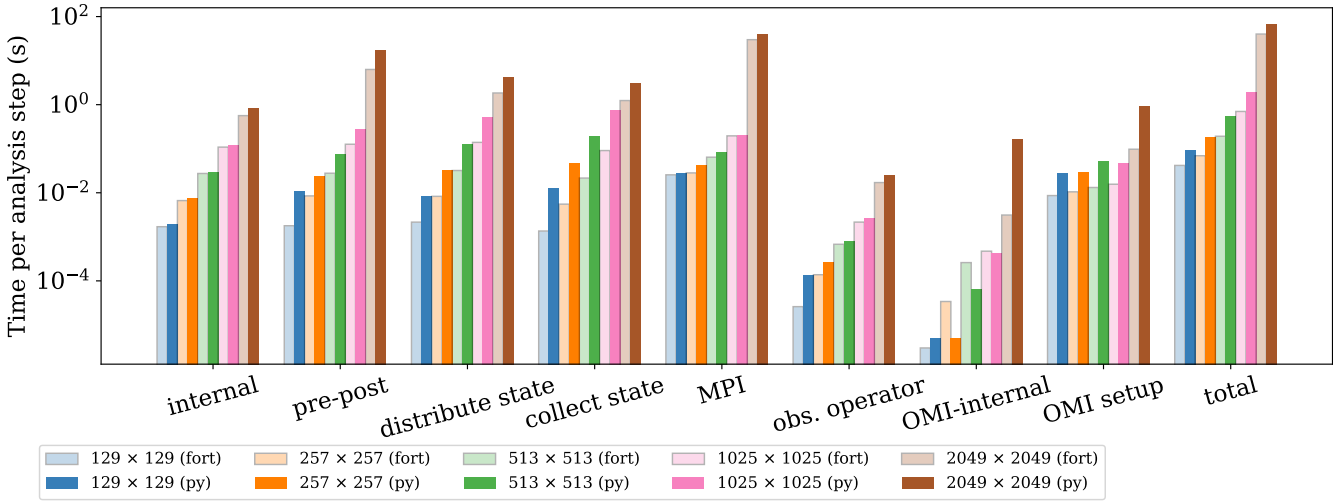

Figure 8. Wall clock time of pyPDAF (dark colour bars) and PDAF (light colour bars) systems per analysis step broken down by functionalities in SCDA ETKF experiments and their total wallclock time per analysis step in log-scale.

The pre- and post-processing is implemented as a user-supplied function (see Sect. 3.3) which is computationally intensive. The intensive computations suit well for the use of the Python JIT compilation. The computational time of the pre- and post-processing increases with the size of the state vector, and Python is in general slower than the Fortran implementation. The difference of wall clock time between the pyPDAF and PDAF-based DA system decreases with increasing state vector size as the overhead in pyPDAF becomes less significant compared to the floating-point computations. As a comparison, on a 129×129 grid, the PDAF system takes 0.04 seconds while the pyPDAF system takes 0.09 seconds per analysis, thus a factor of 2.15 longer time. However, on a 2049×2049 grid, the PDAF system takes around 40.09 seconds per analysis step while the pyPDAF system takes 67.96 seconds per analysis step, thus a factor of only 1.7 longer time. The overhead in pyPDAF system is also comparatively small in high-dimensional systems for the distribution and collection of state vector (labeled 'distribute state' and 'collect state'). For example, the pyPDAF system takes a factor of 2.9 more computational time than the PDAF system on a 129×129 grid but only a factor of 1.3 more time is taken by the pyPDAF system than the PDAF system. The overhead in these functions is proportional to the ensemble size as they are called by each ensemble member respectively. In addition to assigning a state vector to model fields and vice versa in Python, these user-supplied functions perform conversion between physical and spectral space based on Eq. (1) and (2). The transformation utilises the same Fortran subroutines for both PDAF and pyPDAF system. In the pyPDAF system, the Fortran subroutines are converted to Python functions by 'f2py'. The computational time taken by these functions is proportional to the number of grid points. The MPI communications are internal to PDAF which show little differences between pyPDAF and PDAF system.

The wall clock time used for handling observations shows that a pyPDAF DA system is in general slower than a PDAF system. With low-dimensional state vector, the observation operator (labeled 'obs. operator') is slower in a pyPDAF system than PDAF even if the observation operator function only calls a PDAF subroutine provided by OMI. The slow-down of the

pyPDAF system is again a result of overhead in the conversion of Fortran and Python arrays. Here, similar to the collection and distribution of the state vector, the function is called by each ensemble member. The overhead becomes less significant for high-dimensional state vectors when the observation operator computation dominates the total computational time. The internal operations of OMI (labeled 'OMI internal') are very efficient and the pyPDAF systems can be more efficient than PDAF systems. Our experiments do not show clear benefits between pyPDAF and PDAF system for these operations, as expected. The setup of the OMI functionality is implemented in the user-supplied function of *init_dim_obs* (see Sect. 3.3). This includes reading and processing the observation data and their errors. In this case, the pyPDAF-based system is more expensive than the PDAF system. The pyPDAF system is 2.15 (8.57) times slower in executing *init_dim_obs* than the PDAF system on a 129×129 (2049×2049) grid. The relative increase is due to a larger number of observations that needs to be processed.

Our comparison shows that the interfacing between Python and Fortran yields overheads in pyPDAF system even if we utilise JIT compilation of Python. Users need to consider a trade-off between these overheads and the ease of implementation in pyPDAF compared to PDAF. The differences of the computational cost can be less significant for high-dimensional systems for ETKF DA system without localisation.

In practice, localisation is used to avoid sampling errors in high-dimensional weather and climate systems. To make full use of the computational resources, these high-dimensional systems are parallelised by domain decomposition. PDAF exploits the feature of these models for domain localisation where the state vector is also domain decomposed. Here, we choose a domain with 257×257 grid points to assess the LETKF with a localisation radius of 1 spatial unit. As no domain decomposition is implemented for MAOOAM, each processor contains $257 \times 257 \times 4$ local domains which is similar to the number of local domains used in a single processor of a domain decomposed global climate model.

For each local domain, the LETKF computes an analysis using observations with a localisation cut-off radius. Hence, the computational cost depends on the observation density. To investigate the effect of increased intensity of computations on the pyPDAF overhead, we add experiments that observe every 4 grid points.

As shown in Fig. 9, the increased observation density leads to an increase in computational time for the internal operations, observation operator, and the OMI-internal operations due to the larger number of locally assimilated observations. The increased observation density shows little influence on the computational cost of other user-supplied functions. However, as the increased observation density leads to more intensive computations, this mitigates the gap of the total computational time between pyPDAF and PDAF system. In particular, the run times for the internal operations of PDAF (not shown) and OMI ('OMI-internal') dominate the overall run time of the analysis step and show little difference for the pyPDAF and PDAF DA systems.

We notice significant overhead in the pyPDAF system for user-supplied functions related to domain localisation. The increased computational time when the number of domains is specified (labeled 'no. domains') is still of an order of 10^{-4} per analysis step which is negligible. The computation is 5.65 times slower in pyPDAF than the PDAF system for the function specifying the dimension of the local state vector ('init local domain'). The increased computational cost is a result of repeated execution of the user-supplied functions for each local domain. Specifically, in our experiment, this user-supplied function is

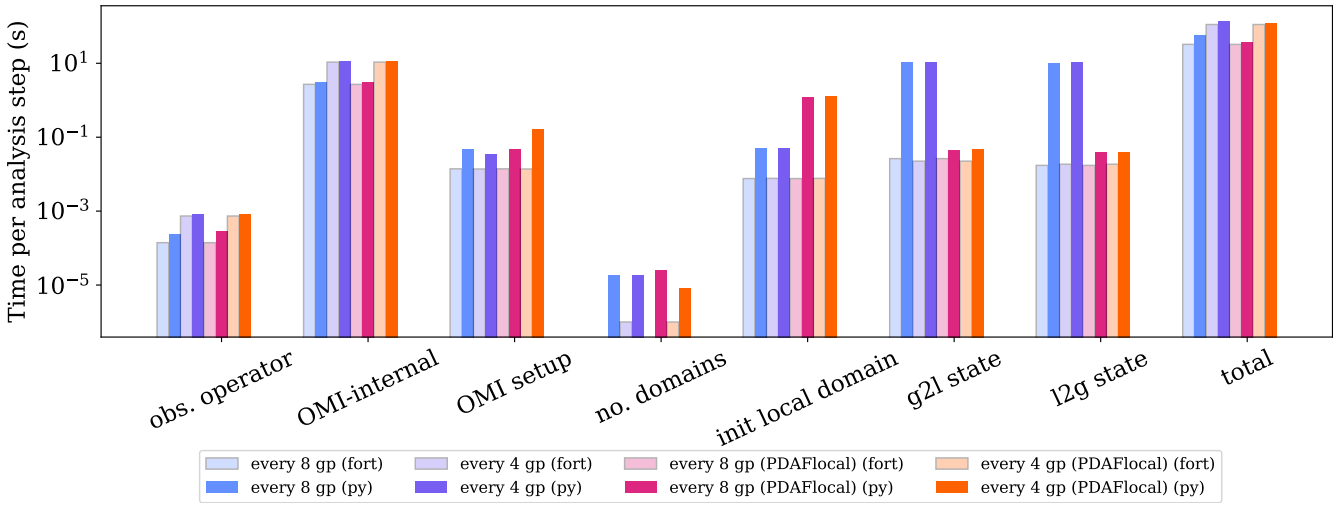

Figure 9. Wall clock time of pyPDAF (light colour bars) and PDAF (dark colour bars) system per analysis step broken down by functionalities in SCDA LETKF experiments and their total wallclock time per analysis step in log-scale. The left four bars (blue and purple bars) represent the case without using the PDAFlocal module while the rest uses the PDAFlocal module. For the sake of conciseness, the functionalities shared by both ETKF and LETKF are omitted.

used $257 \times 257 \times 4$ times per analysis step. The overhead is even higher for the user-supplied functions that convert between local state vector and global state vector ('g2l state' and 'l2g state'), which are called for each ensemble member, due to the conversion of arrays instead of integers. In this experiment, the execution of these routines in pyPDAF system is around 400 times slower than the PDAF system. As these operations are not computationally intensive, the overhead cannot be mitigated by JIT compilation. Without modifications in the PDAF workflow, the overhead can become comparatively less significant with high observation density arising from increased computational cost of other routines, or increased parallelisation of model domains leading to reduced number of local domains on each processor.

To overcome this run time issue of 'g2l state' and 'l2g state', we developed a $PDAFlocal$ module in PDAF, included in release version 2.3, where the user-supplied functions of 'g2l state' and 'l2g state' are circumvented in the PDAF interface as their operations are performed in the compiled Fortran code of $PDAFlocal$. This leads to similar computational cost of these functions between pyPDAF and PDAF system. With $PDAFlocal$, users need to implement an index vector providing the relationship between the state vector in the current local domain and the global state vector when local domain is initialised. Due to this, with PDAFlocal, we see an increased computational time in 'init local domain' in pyPDAF which is around 150 times slower than the PDAF system. The pyPDAF overhead for 'init local domain' is smaller than that of 'g2l state' and 'l2g state' (around 400 times slowdown) due to reduced number of array conversions between Fortran and Python. Further, only one instead of three user-supplied functions are implemented in Python. Due to this, the total computing time is nearly equal for pyPDAF and PDAF with only $6\% - 13\%$ higher time for pyPDAF.

These results demonstrate that pyPDAF can be used with high-dimensional systems with slightly increased overhead per analysis step.

6 Conclusions

We introduce the Python package pyPDAF, which provides an interface to the Parallel Data Assimilation Framework (PDAF). We outline its implementation and design. pyPDAF allows for a Python-based DA system for models coded in Python or interfaced to Python. Furthermore it allows for the implementation of a Python-based offline DA system where the DA is performed separately from the model and data is exchanged between the model and DA code through files. The pyPDAF package allows one to implement user-supplied functions in Python for flexible code development while the DA system still benefits from PDAF's efficient DA algorithm implementation in Fortran.

Using a CDA setup, we demonstrate that pyPDAF can be used with the Python model MAOOAM. Both strongly coupled data assimilation (SCDA) and weakly coupled data assimilation (WCDA) are demonstrated. Our results confirm that the SCDA performs better than WCDA, and additional observations from other model components can improve the overall performance of DA using SCDA. We also investigate the scenario where only one model component is observed. In this case, the error cross-covariance matrix from the ETKF is sufficiently reliable for updating the unobserved model variables leading to improved analyses states for both observed and un-observed model variables. We also show that the DA can improve the long-term trend of the model state in the MAOOAM model.

Using the SCDA setup, the computational costs of using pyPDAF and a Fortran-only implementation with PDAF are compared. We show that the computational time stays the same for the core DA algorithm executed in PDAF while pyPDAF yields an overhead in user-supplied functions. This overhead is a result of both the Python implementation and the requirement of data conversion between Python and Fortran. These overheads become comparatively less significant when the analysis becomes computationally more intensive with increased spatial resolution or observation density. To mitigate the overhead in domain localisation implementations, we introduce a new "PDAFlocal" module in PDAF such that a DA system using pyPDAF can achieve similar computational cost as a pure Fortran based system. This module is included in the release v2.3 of PDAF. We note that JIT compilation or 'f2py' can be used with the Python user-supplied functions for computationally intensive tasks to speed up the Python DA system. Our benchmark shows that, with a global filter, 70% more time is used, and with a domain localised filter, $6\% - 13\%$ more time is used when applying the Python DA system build with pyPDAF in high-dimensional dynamical systems.

pyPDAF opens the possibility to apply sophisticated efficient parallel ensemble DA to large-scale Python models such as machine learning models. pyPDAF also allows for the construction of efficient offline Python DA systems. In particular, pyPDAF can be integrated to machine learning models as long as the state vector can be converted to numpy arrays. A pyPDAF-based DA system allows users to utilise sophisticated parallel ensemble DA methods. However, a pyPDAF system does not support GPU parallelisation like TorchDA (Cheng et al., 2025), which is designed based on the machine learning framework

pyTorch. The TorchDA package may also have limitation on the application of DA on machine learning models implemented by other frameworks.

Code availability. The Fortran and Python code and corresponding configuration and plotting scripts including the randomly generated initial condition for the coupled DA experiments are available at: https://doi.org/10.5281/zenodo.11367123. The MAOOAM V1.4 model used for our experiments is available at https://github.com/Climdyn/MAOOAM/releases/tag/v1.4 with a version available at https://doi.org/10.5281/zenodo.1308192. The Fortran version of the experiment depends on PDAF V2.3 which is released at https://doi.org/10.5281/zenodo.13789628 and can be also found at https://github.com/PDAF/PDAF/releases/tag/PDAF_V2.3 (Nerger, 2024). The source code of pyPDAF is available at https://github.com/yumengch/pyPDAF/releases/tag/v1.0.0 with the exactly same version at https://doi.org/10.5281/zenodo.10950130.

Author contributions. YC coded and distributed the pyPDAF package, conducted the experiments, performed the data analysis, and wrote the paper. LN coded the PDAFlocal module. All authors contribute to the conceptual experiment design and the paper writing.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. The authors acknowledge the UK National Environment Research Council's support for the National Centre for Earth Observation (Contract Number: PR140015).

References

Abernathey, R., rochanotes, Ross, A., Jansen, M., Li, Z., Poulin, F. J., Constantinou, N. C., Sinha, A., Balwada, D., SalahKouhen, Jones, S., Rocha, C. B., Wolfe, C. L. P., Meng, C., van Kemenade, H., Bourbeau, J., Penn, J., Busecke, J., Bueti, M., and Tobias: pyqg/pyqg: v0.7.2, Zenodo [code], https://doi.org/10.5281/zenodo.6563667, 2022.

Ahmed, S. E., Pawar, S., and San, O.: PyDA: A Hands-On Introduction to Dynamical Data Assimilation with Python, Fluids, 5, https://doi.org/10.3390/fluids5040225, 2020.

Anderson, J., Hoar, T., Raeder, K., Liu, H., Collins, N., Torn, R., and Avellano, A.: The Data Assimilation Research Testbed: A Community Facility, Bulletin of the American Meteorological Society, 90, 1283 – 1296, https://doi.org/https://doi.org/10.1175/2009BAMS2618.1, 2009.

Bannister, R. N.: A review of operational methods of variational and ensemble-variational data assimilation, Quarterly Journal of the Royal Meteorological Society, 143, 607–633, https://doi.org/https://doi.org/10.1002/qj.2982, 2017.

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., and Tian, Q.: Accurate medium-range global weather forecasting with 3D neural networks, Nature, 619, 533–538, https://doi.org/10.1038/s41586-023-06185-3, 2023.

Bishop, C. H., Etherton, B. J., and Majumdar, S. J.: Adaptive Sampling with the Ensemble Transform Kalman Filter. Part I: Theoretical Aspects, Monthly Weather Review, 129, 420 – 436, https://doi.org/https://doi.org/10.1175/1520-0493(2001)129<0420:ASWTET>2.0.CO;2, 2001.

Bonavita, M., Hólm, E., Isaksen, L., and Fisher, M.: The evolution of the ECMWF hybrid data assimilation system, Quarterly Journal of the Royal Meteorological Society, 142, 287–303, https://doi.org/https://doi.org/10.1002/qj.2652, 2016.

Bruggeman, J., Bolding, K., Nerger, L., Teruzzi, A., Spada, S., Skákala, J., and Ciavatta, S.: EAT v1.0.0: a 1D test bed for physical–biogeochemical data assimilation in natural waters, Geoscientific Model Development, 17, 5619–5639, https://doi.org/10.5194/gmd-17-5619-2024, 2024.

Caron, J.-F., Milewski, T., Buehner, M., Fillion, L., Reszka, M., Macpherson, S., and St-James, J.: Implementation of Deterministic Weather Forecasting Systems Based on Ensemble–Variational Data Assimilation at Environment Canada. Part II: The Regional System, Monthly Weather Review, 143, 2560 – 2580, https://doi.org/10.1175/MWR-D-14-00353.1, 2015.

Carrassi, A., Bocquet, M., Bertino, L., and Evensen, G.: Data assimilation in the geosciences: An overview of methods, issues, and perspectives, WIREs Climate Change, 9, e535, https://doi.org/https://doi.org/10.1002/wcc.535, 2018.

Cehelsky, P. and Tung, K. K.: Theories of Multiple Equilibria and Weather Regimes—A Critical Reexamination. Part II: Baroclinic Two-Layer Models, Journal of Atmospheric Sciences, 44, 3282 – 3303, https://doi.org/10.1175/1520-0469(1987)044<3282:TOMEAW>2.0.CO;2, 1987.

Cheng, S., Min, J., Liu, C., and Arcucci, R.: TorchDA: A Python package for performing data assimilation with deep learning forward and transformation functions, Computer Physics Communications, 306, 109 359, https://doi.org/https://doi.org/10.1016/j.cpc.2024.109359, 2025.

Clayton, A. M., Lorenc, A. C., and Barker, D. M.: Operational implementation of a hybrid ensemble/4D-Var global data assimilation system at the Met Office, Quarterly Journal of the Royal Meteorological Society, 139, 1445–1461, https://doi.org/https://doi.org/10.1002/qj.2054, 2013.

De Cruz, L., Demaeyer, J., and Vannitsem, S.: The Modular Arbitrary-Order Ocean-Atmosphere Model: MAOOAM v1.0, Geoscientific Model Development, 9, 2793–2808, https://doi.org/10.5194/gmd-9-2793-2016, 2016.

de Rosnay, P., Browne, P., de Boisséson, E., Fairbairn, D., Hirahara, Y., Ochi, K., Schepers, D., Weston, P., Zuo, H., Alonso-Balmaseda, M., Balsamo, G., Bonavita, M., Borman, N., Brown, A., Chrust, M., Dahoui, M., Chiara, G., English, S., Geer, A., Healy, S., Hersbach, H., Laloyaux, P., Magnusson, L., Massart, S., McNally, A., Pappenberger, F., and Rabier, F.: Coupled data assimilation at ECMWF: current status, challenges and future developments, Quarterly Journal of the Royal Meteorological Society, 148, 2672–2702, https://doi.org/https://doi.org/10.1002/qj.4330, 2022.

Döll, P., Hasan, H. M. M., Schulze, K., Gerdener, H., Börger, L., Shadkam, S., Ackermann, S., Hosseini-Moghari, S.-M., Müller Schmied, H., Güntner, A., and Kusche, J.: Leveraging multi-variable observations to reduce and quantify the output uncertainty of a global hydrological model: evaluation of three ensemble-based approaches for the Mississippi River basin, Hydrology and Earth System Sciences, 28, 2259–2295, https://doi.org/10.5194/hess-28-2259-2024, 2024.

Evensen, G.: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics, Journal of Geophysical Research: Oceans, 99, 10 143–10 162, https://doi.org/https://doi.org/10.1029/94JC00572, 1994.

Evensen, G., Vossepoel, F. C., and van Leeuwen, P. J.: Data assimilation fundamentals: A unified formulation of the state and parameter estimation problem, Springer Nature, 2022.

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., and Taylor, K. E.: Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization, Geoscientific Model Development, 9, 1937–1958, https://doi.org/10.5194/gmd-9-1937-2016, 2016.

Feng, L., Palmer, P., Bösch, H., and Dance, S.: Estimating surface CO 2 fluxes from space-borne CO 2 dry air mole fraction observations using an ensemble Kalman Filter, Atmospheric chemistry and physics, 9, 2619–2633, 2009.

filterpy PyPI: https://pypi.org/project/filterpy/, last access: 2024-08-29.

Hamill, T. M., Whitaker, J. S., and Snyder, C.: Distance-Dependent Filtering of Background Error Covariance Estimates in an Ensemble Kalman Filter, Monthly Weather Review, 129, 2776 – 2790, https://doi.org/10.1175/1520-0493(2001)129<2776:DDFOBE>2.0.CO;2, 2001.

Hamill, T. M., Whitaker, J. S., Fiorino, M., and Benjamin, S. G.: Global Ensemble Predictions of 2009's Tropical Cyclones Initialized with an Ensemble Kalman Filter, Monthly Weather Review, 139, 668 – 688, https://doi.org/10.1175/2010MWR3456.1, 2011.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.-N.: The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society, 146, 1999–2049, https://doi.org/https://doi.org/10.1002/qj.3803, 2020.

Houtekamer, P. L., Mitchell, H. L., Pellerin, G., Buehner, M., Charron, M., Spacek, L., and Hansen, B.: Atmospheric Data Assimilation with an Ensemble Kalman Filter: Results with Real Observations, Monthly Weather Review, 133, 604 – 620, https://doi.org/10.1175/MWR-2864.1, 2005.

Hu, C.-C. and van Leeuwen, P. J.: A particle flow filter for high-dimensional system applications, Quarterly Journal of the Royal Meteorological Society, 147, 2352–2374, https://doi.org/https://doi.org/10.1002/qj.4028, 2021.

Hunt, B. R., Kostelich, E. J., and Szunyogh, I.: Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter, Physica D: Nonlinear Phenomena, 230, 112–126, https://doi.org/https://doi.org/10.1016/j.physd.2006.11.008, 2007.

Kalnay, E., Sluka, T., Yoshida, T., Da, C., and Mote, S.: Review article: Towards strongly coupled ensemble data assimilation with additional improvements from machine learning, Nonlinear Processes in Geophysics, 30, 217–236, https://doi.org/10.5194/npg-30-217-2023, 2023.

Kurth, T., Subramanian, S., Harrington, P., Pathak, J., Mardani, M., Hall, D., Miele, A., Kashinath, K., and Anandkumar, A.: FourCastNet: Accelerating Global High-Resolution Weather Forecasting Using Adaptive Fourier Neural Operators, in: The Platform for Advanced Scientific Computing 2023, Association for Computing Machinery, New York, NY, USA, https://doi.org/10.1145/3592979.3593412, 2023.

Kurtz, W., He, G., Kollet, S. J., Maxwell, R. M., Vereecken, H., and Hendricks Franssen, H.-J.: TerrSysMP–PDAF (version 1.0): a modular high-performance data assimilation framework for an integrated land surface–subsurface model, Geoscientific Model Development, 9, 1341–1360, https://doi.org/10.5194/gmd-9-1341-2016, 2016.

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., Merose, A., Hoyer, S., Holland, G., Vinyals, O., Stott, J., Pritzel, A., Mohamed, S., and Battaglia, P.: Learning skillful medium-range global weather forecasting, Science, 382, 1416–1421, https://doi.org/10.1126/science.adi2336, 2023.

Losa, S. N., Danilov, S., Schröter, J., Nerger, L., Maβmann, S., and Janssen, F.: Assimilating NOAA SST data into the BSH operational circulation model for the North and Baltic Seas: Inference about the data, Journal of Marine Systems, 105-108, 152–162, https://doi.org/https://doi.org/10.1016/j.jmarsys.2012.07.008, 2012.

McGibbon, J., Brenowitz, N. D., Cheeseman, M., Clark, S. K., Dahm, J. P. S., Davis, E. C., Elbert, O. D., George, R. C., Harris, L. M., Henn, B., Kwa, A., Perkins, W. A., Watt-Meyer, O., Wicky, T. F., Bretherton, C. S., and Fuhrer, O.: fv3gfs-wrapper: a Python wrapper of the FV3GFS atmospheric model, Geoscientific Model Development, 14, 4401–4409, https://doi.org/10.5194/gmd-14-4401-2021, 2021.

Message Passing Interface Forum: MPI: A Message-Passing Interface Standard Version 4.1, https://www.mpi-forum.org/docs/mpi-4.1/mpi41-report.pdf, 2023.

Nerger, L.: Data assimilation for nonlinear systems with a hybrid nonlinear Kalman ensemble transform filter, Quarterly Journal of the Royal Meteorological Society, 148, 620–640, https://doi.org/https://doi.org/10.1002/qj.4221, 2022.

Nerger, L.: PDAF Version 2.3, Zenodo [code], https://doi.org/10.5281/zenodo.13789628, 2024.

Nerger, L. and Hiller, W.: Software for ensemble-based data assimilation systems—Implementation strategies and scalability, Computers & Geosciences, 55, 110–118, https://doi.org/https://doi.org/10.1016/j.cageo.2012.03.026, ensemble Kalman filter for data assimilation, 2013a.

Nerger, L. and Hiller, W.: Software for Ensemble-based Data Assimilation Systems - Implementation Strategies and Scalability, Computers & Geosciences, 55, 110–118, 2013b.

Nerger, L., Hiller, W., and Schröter, J.: PDAF - The parallel data assimilation framework: experiences with Kalman filtering, in: Use of High Performance Computing in Meteorology, pp. 63–83, https://doi.org/10.1142/9789812701831_0006, 2005.

Nerger, L., Janjić, T., Schröter, J., and Hiller, W.: A Unification of Ensemble Square Root Kalman Filters, Monthly Weather Review, 140, 2335 – 2345, https://doi.org/https://doi.org/10.1175/MWR-D-11-00102.1, 2012.

Nerger, L., Tang, Q., and Mu, L.: Efficient ensemble data assimilation for coupled models with the Parallel Data Assimilation Framework: example of AWI-CM (AWI-CM-PDAF 1.0), Geoscientific Model Development, 13, 4305–4321, https://doi.org/10.5194/gmd-13-4305-2020, 2020.

PDAF - the Parallel Data Assimilation Framework: https://pdaf.awi.de/, last access: 2024-02-13.

Penny, S. G. and Hamill, T. M.: Coupled data assimilation for integrated earth system analysis and prediction, Bulletin of the American Meteorological Society, 98, ES169–ES172, 2017.

Pham, D. T.: Stochastic Methods for Sequential Data Assimilation in Strongly Nonlinear Systems, Monthly Weather Review, 129, 1194 – 1207, https://doi.org/https://doi.org/10.1175/1520-0493(2001)129<1194:SMFSDA>2.0.CO;2, 2001.

Pham, D. T., Verron, J., and Christine Roubaud, M.: A singular evolutive extended Kalman filter for data assimilation in oceanography, Journal of Marine Systems, 16, 323–340, https://doi.org/https://doi.org/10.1016/S0924-7963(97)00109-7, 1998.

Pohlmann, H., Brune, S., Fröhlich, K., Jungclaus, J. H., Sgoff, C., and Baehr, J.: Impact of ocean data assimilation on climate predictions with ICON-ESM, Climate Dynamics, 61, 357–373, https://doi.org/10.1007/s00382-022-06558-w, 2023.

Raanes, P. N., Chen, Y., and Grudzien, C.: DAPPER: Data Assimilation with Python: a Package for Experimental Research, Journal of Open Source Software, 9, 5150, https://doi.org/10.21105/joss.05150, 2024.

Sakov, P. and Oke, P. R.: A deterministic formulation of the ensemble Kalman filter: an alternative to ensemble square root filters, Tellus A, 60, 361–371, https://doi.org/10.1111/j.1600-0870.2007.00299.x, 2008.

Sakov, P., Counillon, F., Bertino, L., Lisæter, K. A., Oke, P. R., and Korablev, A.: TOPAZ4: an ocean-sea ice data assimilation system for the North Atlantic and Arctic, Ocean Science, 8, 633–656, https://doi.org/10.5194/os-8-633-2012, 2012.

SALOME The Open Source Integration Platform for Numerical Simulation: http://www.salome-platform.org/, last access: 2024-08-29.

Shao, C. and Nerger, L.: WRF-PDAF v1.0: implementation and application of an online localized ensemble data assimilation framework, Geoscientific Model Development, 17, 4433–4445, https://doi.org/10.5194/gmd-17-4433-2024, 2024.

Sluka, T. C., Penny, S. G., Kalnay, E., and Miyoshi, T.: Assimilating atmospheric observations into the ocean using strongly coupled ensemble data assimilation, Geophysical Research Letters, 43, 752–759, https://doi.org/https://doi.org/10.1002/2015GL067238, 2016.

Smith, P. J., Fowler, A. M., and Lawless, A. S.: Exploring strategies for coupled 4D-Var data assimilation using an idealised atmosphere–ocean model, Tellus A: Dynamic Meteorology and Oceanography, https://doi.org/10.3402/tellusa.v67.27025, 2015.

Strebel, L., Bogena, H. R., Vereecken, H., and Hendricks Franssen, H.-J.: Coupling the Community Land Model version 5.0 to the parallel data assimilation framework PDAF: description and applications, Geoscientific Model Development, 15, 395–411, https://doi.org/10.5194/gmd-15-395-2022, 2022.

Tang, Q., Mu, L., Goessling, H. F., Semmler, T., and Nerger, L.: Strongly coupled data assimilation of ocean observations into an ocean-atmosphere model, Geophys. Res. Lett., 48, e2021GL094 941, 2021.

Tang, Q., Delottier, H., Kurtz, W., Nerger, L., Schilling, O. S., and Brunner, P.: HGS-PDAF (version 1.0): a modular data assimilation framework for an integrated surface and subsurface hydrological model, Geoscientific Model Development, 17, 3559–3578, https://doi.org/10.5194/gmd-17-3559-2024, 2024.

The Python Language Reference: https://docs.python.org/3/reference/introduction.html#alternate-implementations, last access: 2024-02-13.

Tondeur, M., Carrassi, A., Vannitsem, S., and Bocquet, M.: On temporal scale separation in coupled data assimilation with the ensemble kalman filter, Journal of Statistical Physics, 179, 1161–1185, https://doi.org/10.1007/s10955-020-02525-z, 2020.

Trémolet, Y. and Auligne, T.: The Joint Effort for Data Assimilation Integration (JEDI), JCSDA Q, 66, 1–5, 2020.

Tödter, J. and Ahrens, B.: A Second-Order Exact Ensemble Square Root Filter for Nonlinear Data Assimilation, Monthly Weather Review, 143, 1347 – 1367, https://doi.org/10.1175/MWR-D-14-00108.1, 2015.

van Leeuwen, P. J., Künsch, H. R., Nerger, L., Potthast, R., and Reich, S.: Particle filters for high-dimensional geoscience applications: A review, Quarterly Journal of the Royal Meteorological Society, 145, 2335–2365, https://doi.org/https://doi.org/10.1002/qj.3551, 2019.

Vetra-Carvalho, S., van Leeuwen, P. J., Nerger, L., Barth, A., Altaf, M. U., Brasseur, P., Kirchgessner, P., and Beckers, J.-M.: State-of-the-art stochastic data assimilation methods for high-dimensional non-Gaussian problems, Tellus A, 70, 1445 364, 2018.

Villa, U., Petra, N., and Ghattas, O.: HIPPYlib: An Extensible Software Framework for Large-Scale Inverse Problems Governed by PDEs: Part I: Deterministic Inversion and Linearized Bayesian Inference, ACM Trans. Math. Softw., 47, https://doi.org/10.1145/3428447, 2021.

Walters, D., Baran, A. J., Boutle, I., Brooks, M., Earnshaw, P., Edwards, J., Furtado, K., Hill, P., Lock, A., Manners, J., Morcrette, C., Mulcahy, J., Sanchez, C., Smith, C., Stratton, R., Tennant, W., Tomassini, L., Van Weverberg, K., Vosper, S., Willett, M., Browse, J., Bushell, A., Carslaw, K., Dalvi, M., Essery, R., Gedney, N., Hardiman, S., Johnson, B., Johnson, C., Jones, A., Jones, C., Mann, G., Milton, S., Rumbold, H., Sellar, A., Ujiie, M., Whitall, M., Williams, K., and Zerroukat, M.: The Met Office Unified Model Global Atmosphere 7.0/7.1 and JULES Global Land 7.0 configurations, Geoscientific Model Development, 12, 1909–1963, https://doi.org/10.5194/gmd-12-1909-2019, 2019.

Williams, N., Byrne, N., Feltham, D., Van Leeuwen, P. J., Bannister, R., Schroeder, D., Ridout, A., and Nerger, L.: The effects of assimilating a sub-grid-scale sea ice thickness distribution in a new Arctic sea ice data assimilation system, The Cryosphere, 17, 2509–2532, https://doi.org/10.5194/tc-17-2509-2023, 2023.

Ying, Y. M.: nansencenter/NEDAS: v1.0-beta, Zenodo [code], https://doi.org/10.5281/zenodo.10525331, 2024.

Zhang, S., Liu, Z., Zhang, X., Wu, X., Han, G., Zhao, Y., Yu, X., Liu, C., Liu, Y., Wu, S., et al.: Coupled data assimilation and parameter estimation in coupled ocean–atmosphere models: a review, Climate Dynamics, 54, 5127–5144, https://doi.org/https://doi.org/10.1007/s00382-020-05275-6, 2020.

Zhao, F., Liang, X., Tian, Z., Li, M., Liu, N., and Liu, C.: Southern Ocean Ice Prediction System version 1.0 (SOIPS v1.0): description of the system and evaluation of synoptic-scale sea ice forecasts, Geoscientific Model Development, 17, 6867–6886, https://doi.org/10.5194/gmd-17-6867-2024, 2024.

Zhu, M., van Leeuwen, P. J., and Amezcua, J.: Implicit equal-weights particle filter, Quarterly Journal of the Royal Meteorological Society, 142, 1904–1919, https://doi.org/https://doi.org/10.1002/qj.2784, 2016.