https://doi.org/10.5194/egusphere-2022-114
Preprint. Discussion started: 20 May 2022

Development of a flexible data assimilation method in a 3D unstructured-grid ocean model under Earth System Modeling Framework

Hao-Cheng Yu[1], Y. Joseph Zhang[1], Lars Nerger[2], Carsten Lemmen[3], Jason C.S. Yu[4], Tzu-Yin Chou[4], Chi-Hao Chu[5], Chuen-Teyr Terng[5]

[1]Center for Coastal Resource Management, Virginia Institute of Marine Science, College of William & Mary, Gloucester Point, 23062, USA
[2]Alfred-Wegener-Institut Helmholtz-Zentrum für Polar und Meerresforschung, Bremerhaven, 27570, Germany
[3]Institute of Coastal Systems Analysis and Modeling, Helmholtz-Zentrum Hereon, Geesthacht, 21502, Germany
[4]Department of Marine Environment and Engineering, National Sun Yat-Sen University, Kaohsiung, 80424, Taiwan
[5]Marine Meteorology Center, Central Weather Bureau, Taipei, 100006, Taiwan

Correspondence to: Hao-Cheng Yu (hcyu@vims.edu) and Y. Joseph Zhang (yjzhang@vims.edu)

Abstract. We develop a new data assimilative (DA) approach by combining two parallel frameworks: a parallel DA framework (PDAF) and a flexible model coupling framework (ESMF). The new DA system is built on the ESMF at the top level that drives the PDAF and any combination of Earth system modeling (ESM) components, to allow maximum flexibility and easy implementation of data assimilation for fully coupled ESM applications. We demonstrate the new DA system using a 3D unstructured-grid ocean model as ESM in this paper. The new system is validated using a simple benchmark and applied to a realistic case of Kuroshio simulation around Taiwan. The new system is demonstrated to significantly improve the model skill for temperature, velocity and surface elevation before, during and after typhoon events. The flexibility and ease of implementation make the new system widely applicable for other coupled ESMs.

1 Introduction

Ensemble based data assimilation (DA) approaches are popular choices for DA due to their advantage of being less intrusive to the original model code than the adjoint based approaches (Kalnay et al., 2007, Carrassi et al., 2018, Vetra-Carvalho et al., 2018). A prominent example of a community-supported package is the Parallel Data Assimilation Framework (PDAF; Nerger et al. 2005, Nerger and Hiller 2013), which has been successfully utilized by several Earth system models (ESMs) such as FESOM, AWI-CM, NEMO, MITgcm etc. It is model agnostic and can handle different types of ESMs written in modern parallel computing languages, e.g., structured or unstructured grid, explicit or implicit models. It allows the choices of offline and online coupling between the DA component and the ESMs. While the offline coupling mode interfaces between the model and data assimilation codes using disk files and is hence less intrusive to the model code, the online mode

https://doi.org/10.5194/egusphere-2022-114
Preprint. Discussion started: 20 May 2022

augments model code with DA functionality and is significantly more efficient and allows the nonlinear feedback from DA results to the 'forward' model. We will focus on the online mode in this paper.

A large and growing selection of ensemble filters of both ensemble Kalman and nonlinear particle filters is available inside PDAF (Table 1). Furthermore, 3D variational (3DVar) methods (see Bannister, 2017) can also be supported by PDAF; parameterized covariances can be used to represent the uncertainty in the model state or ensembles, which are propagated using the same ensemble infrastructure as the ensemble filters for maximum efficiency. For operations, PDAF supports two modes: (1) fully parallel mode where all ensemble members are executed simultaneously during time stepping; (2) flexible mode where subsets of members ('cohorts' as shown in Fig. 2) are executed in batches. Obviously, the fully parallel mode allows maximum efficiency as measured by 'time-to-solution', but the flexible mode is more practical when computational resources are limited. This flexibility is very important for practical applications because computational resources may not always be sufficient to support the fully parallel mode (Valcke 2022).

Despite its success, the PDAF enabled DA applications have mostly focused on a single ESM component so far, like the ocean (e.g. Nerger et al. 2007, Brune et al, 2015), sea-ice (e.g., Yang et al., 2014, Mu et al., 2019), ocean-biogeochemistry (e.g., Pradhan et al., 2019, 2020; Goodliff et al., 2019), atmospheric transport (Pardini, 2020), or the solid Earth (Fournier et al., 2013; Schachtschneider et al. 2022). While extension to cover multiple coupled components is feasible with PDAF (Nerger et al., 2020, Kurtz et al, 2016), this often involves non-trivial amount of developmental work. The ensemble based DA systems such as PDAF require time-stepping of multiple ensemble members of the forward model, often written in domain decomposition based MPI parallelism (Gropp et al., 1994). While PDAF allows partitioning of global ('world') communication space into sub-communicators for each 'instance' of the model (ensemble member), it is challenging to generalize it to account for complex nonlinear coupling between ESMs in the future. Therefore, we implement the new DA system under the umbrella of a parallel coupling framework, Earth System Modeling Framework (ESMF, https://earthsystemmodeling.org/, last access: 10 May 2022). This allows DA to be performed on fully coupled ESMs running on different grids/meshes as ESMF can perform regridding and interpolation between different ESM components.

In this paper, we develop a DA system based on ESMF-PDAF that can be easily extended to assimilate multiple coupled ESMs simultaneously. To clearly demonstrate the methodology, however, in this paper we will only validate the new DA system with an ocean model. Extension to coupled ESMs (e.g., ocean-atmosphere-biology etc) is trivial. The paper is organized as follows. We first describe the building blocks and overall structure of the new DA system in Section 2. We then validate the new system using a simple twin experiment in Section 3.1, before applying it to a challenging realistic case of northwestern Pacific in Section 3.2. Efficiency and overhead of the new DA system are examined in Section 3.3 using the realistic case. Finally, we summarize the major findings in this paper and future work in Section 4.

https://doi.org/10.5194/egusphere-2022-114
Preprint. Discussion started: 20 May 2022

2 SCHISM-PDAF implementation under ESMF

ESMF is a suite of software tools for developing high-performance, multi-component Earth science modeling applications. Such applications may include many components representing atmospheric, oceanic, terrestrial, or other physical domains, and their constituent processes (dynamical, chemical, biological, etc.). Often these components are developed by different groups independently and must be "coupled" together using software that transfers and transforms data among the components in order to form functional simulations. ESMF supports the development of these complex applications in a number of ways. It introduces a set of simple, consistent component interfaces that apply to all types of components, including couplers themselves. These interfaces expose in an obvious way the inputs and outputs of each component. It offers a variety of data structures for transferring data between components, and libraries for regridding, time advancement, and other common modeling functions. ESMF coupling has been used to construct modular ESMs at low, high, and flexible granularity (e.g. the Community Earth System Model CESM, the Goddard Earth Observing System (GEOS-5, Ott et al. 2009), the Modular System for Shelves and Coasts MOSSCO, Lemmen et al. 2018). In general, using ESMF as mediator to link all components leads to less intrusion to the model code and maintain independence for each library.

Once the parallel world is partitioned into smaller worlds using ESMF for each ensemble member used in the DA (which is executed by a persistent execution thread (PET)), the ESM and PDAF interface codes are then inserted into the main ESMF driver (Fig. 1). The PDAF interface is relatively straightforward and consists of initialization (at the start of simulation), and a few subroutine calls to PDAF during time stepping. The latter include PDAF_get_state() for updating member state and assimilate_pdaf() at the specified times of assimilation. The types of filters available in the latest PDAF version are described in Table 1. Currently we have fully implemented 2 types of filters (ETKF, ESTKF) together with their localization variants (LETKF, LESTKF) with plans for other filters in the near future.

On the ESM (ocean model) side, we first supply PDAF with the required 'binding' (interface) codes to link PDAF and the ocean model, SCHISM (schism.wiki; Zhang et al. 2016). PDAF is designed to link with ESMs model with a more generalized interface through a set of functions (Fortran subroutines in our case). Thus, one can simply follow the online tutorial to link DA capability into any model. Here we list the major interface codes as follows, and more detail can be found in the tutorial (http://pdaf.awi.de/trac/wiki/PdafTutorial, last access: 10 May 2022):

1) collect_state_pdaf/distribute_state_pdaf: create functions to fill model state with specified SCHISM variables or update SCHISM variables with values from state vector;

2) init_dim_obs_pdaf/init_dim_obs_f_pdaf: add interface to read in observation data including error estimates for each type;

3) obs_op_pdaf/obs_op_f_pdaf: choose an observation operator to map model state onto observed state;

4) init_dim_l_pdaf: specify local analysis domain, e.g., a vertical grid column, and localization radius within which observations are utilized by the local filter;

5) init_dim_obs_l_pdaf: locate elements of model state in local analysis domain;

6) l2g_state_pdaf/g2l_state_pdaf: add 'local to global' / 'global to local' model state conversion routine for local filter;

https://doi.org/10.5194/egusphere-2022-114
Preprint. Discussion started: 20 May 2022

7) output_netcdf_pdaf: add interface to output DA (ensemble mean) result.

Furthermore, to accommodate the flexible mode of PDAF, i.e., using a fewer number of PETs than the number of ensemble members, the time stepping part of SCHISM is modified. Significantly, the flexible mode requires 'rewinding' of time clock after an ensemble member reaches the next assimilation time. This is because another ensemble member shares the same memory space, and the model forcing (wind, etc) needs to go back to the starting time to redo the time stepping (Fig. 2). To this end, we have modified the reading of forcing inside SCHISM to allow arbitrary rewinding.

At the lowest level of the DA system is the Earth system model component. In this paper we will focus on a single component, the ocean model. SCHISM is an open-source 3D baroclinic model for cross-scale hydrodynamic and hydrologic applications. It uses unstructured quadrangular-triangular elements in the horizontal dimension and a very flexible vertical gridding system (LSC^2; Zhang et al. 2015). The model has been successfully applied to many nearshore and offshore systems around the world (see schism.wiki for a complete publication list). Related to this study, the model has been used as the engine for the operational forecasting system for Taiwan since 2011 (Yu et al. 2017; https://npd.cwb.gov.tw/NPD/products_display/product?menu_index=4, last access: 10 May 2022).

Extension to coupled ESMs with the current DA system involves preparation of the binding codes (1-7) for each ESM component, and addition of ESM interfaces inside the main ESMF driver similar to the SCHISM part in Fig. 1. The complex model coupling is handled by ESMF in a modular way using high-level function calls.

https://doi.org/10.5194/egusphere-2022-114
Preprint. Discussion started: 20 May 2022

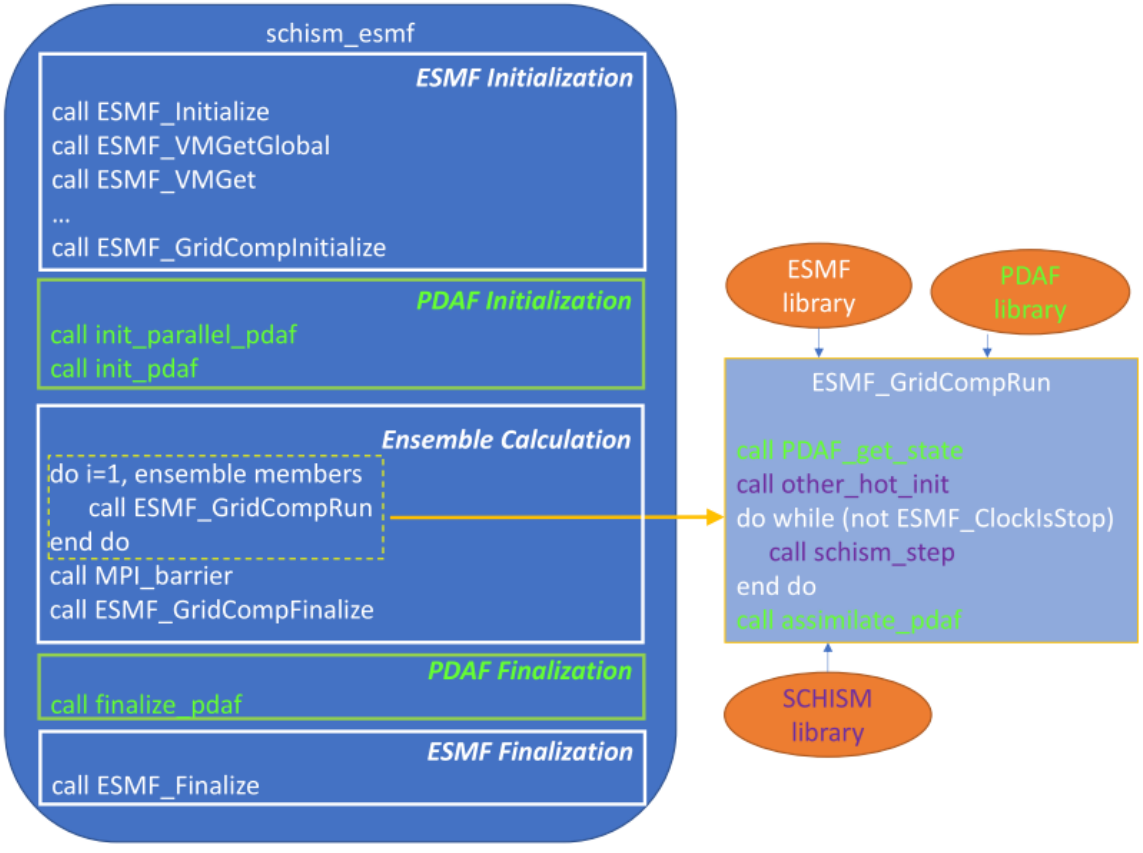

Fig. 1: Schematic code structure for the new DA system. At the top level, ESMF is used to orchestrate the parallel environment and model coupling. It initiates the environment for PDAF (for each ensemble member) and for ESMs. The code requires multiple libraries: ESMF, PDAF and one for each ESM (SCHISM in this case).

https://doi.org/10.5194/egusphere-2022-114
Preprint. Discussion started: 20 May 2022

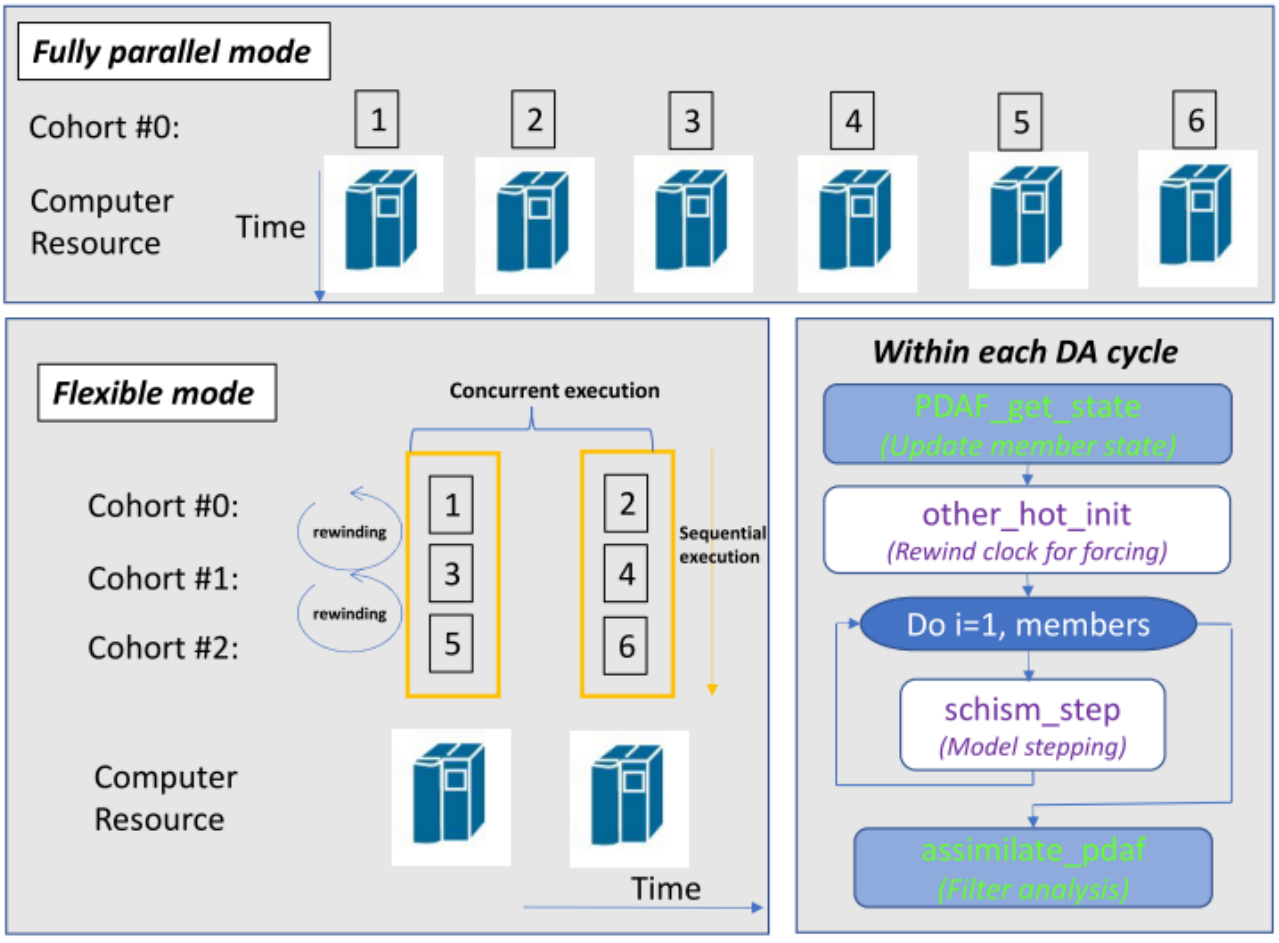

Fig. 2: Time stepping part of the new DA system (i.e., the dashed yellow box in Fig. 1) under fully parallel mode (top panel) and flexible mode (bottom panel). Task IDs (1,2...6) are distributed into 3 'cohorts' for concurrent execution in ESMF under the flexible mode of PDAF. The task IDs in each orange box are executed on the same PET sequentially, and the clock needs to be rewound during the hand-over between '1' and '3' etc via the subroutine 'other_hot_init'.

https://doi.org/10.5194/egusphere-2022-114
Preprint. Discussion started: 20 May 2022

Table 1: Types of filters supported in PDAF

Filter	Localized version	Status	Reference
Error-Subspace Transform Kalman Filter (ESTKF)	LESTKF	Ready	Nerger et al. (2012a, 2012b)
Ensemble Transform Kalman Filter (ETKF)	LETKF	Ready	Hunt et al. (2007)
Ensemble Kalman Filter (EnKF)	LEnKF	In development	Evensen (1994)
Singular Evolutive Extended Kalman (SEEK)		In development	Pham et al. (1998b)
Singular Evolutive Interpolated extended Kalman (SEIK)	LSEIK	In development	Pham et al. (1998a, 2001)
Non-linear Ensemble Transform Kalman Filter (NETF)	LNETF	In development	Tödter & Ahrens (2015)
Particle Filter with resampling (PF)		In development	Vetra-Carvalho Sanita et al. (2018)
3DVar with parameterized covariance matrix (3DVar)		In development	Bannister (2017)
3DVar using ensemble covariance matrix (3DEnVar)	Ensemble perturbations are updated with the LESTKF filter	In development	Bannister (2017)
Hybrid 3DVar using a combination of parameterized and ensemble covariance matrix (Hyb3DVar)	Ensemble perturbations are updated with the LESTKF filter	In development	Bannister (2017)

3 Validation and application of the new DA system

In this section, we will first validate the new DA system using a simple idealized test with manufactured 'observations'. We then apply the system to a realistic and challenging case of simulating typhoons in the northwestern Pacific with a focus on the regions near Taiwan using satellite and in-situ profiler observations. We will also demonstrate the efficiency and overhead of the DA system in the realistic case.

https://doi.org/10.5194/egusphere-2022-114
Preprint. Discussion started: 20 May 2022

3.1 Lock-exchange test

3.1.1 Test description

To verify that this DA framework works as intended, we start from a simple test case with lock-exchange experiment. This test case is initialized with a horizontally varying (but vertically uniform) temperature distribution (10℃ at x=0km and linearly increases to 16℃ at x=20km) in a narrow, closed tank with sloped bathymetry (Fig. 3). The tank dimension is 20km in length with 1km width, with depth varying from 50m to 10m with a linear slope between x=6km and 13 km (Fig. 3). The only external forcing is the surface wind that mixes the water.

3.1.2 Generation of ensemble and observation

The generation of an initial ensemble of model state realizations is a key step in PDAF and a good ensemble spread is desirable to cover potential model state trajectory to ensure success of assimilation (Vetra-Carvalho et al., 2018). Here we follow Pham's method (Pham et al. 2001) to generate member states as suggested by PDAF. We first conduct a 5-day free model run and extract hourly snapshots of the model state during the entire simulation period. With EOF-decomposition analysis, singular vectors and values are extracted from these snapshots. The ensemble member states are then initialized from these values; a particular matrix that represents the model mean and covariance is generated by multiplying each singular vector with the corresponding singular value. This matrix is then multiplied by a random matrix with properties that ensure conservation of the mean state and covariances to yield the ensemble perturbations. This method basically utilizes the inherent variability of model dynamics to represent uncertainty.

We follow the common twin-experiment setup here. The free run snapshots are used as "truth". Observation data is then extracted from the free run by adding Gaussian-distributed random values within a standard deviation of 0.15℃ that represents the observation error. To assess how well the assimilation can rectify the 'erroneous' model states, the ensemble model mean is initialized by intentionally adding -1℃ to the first 24-hour mean state.

Major DA parameters include filter type, frequency of DA cycle, forgetting factor, localization parameters and observation error (Carrassi et al. 2018). The observation error, whose estimates usually come with observation data, is found to be an important control. For practical applications, we have implemented 3 types of observation error input options. The first option sets the errors uniformly for all observation types. The second option allows different errors to be used for different variables, such as elevation, temperature etc. The third option is the most flexible, allowing the errors to be specified at each observation point, and is most useful for operational forecasts.

3.1.3 Results from assimilating a single profile

To clearly see how DA affects the model state, we start with only assimilating one near-surface temperature profile (down to 10m). LESTKF filter is used here with 8 ensemble members and 500m as the localization range to perform the analysis. The local filter usually gives better results than the corresponding global filter (cf. Section 3.1.4, Nerger et al. 2006). Fig. 3 shows

https://doi.org/10.5194/egusphere-2022-114
Preprint. Discussion started: 20 May 2022

the difference between before and after assimilation. The 'forecast' is the ensemble mean just before the DA, while the 'analysis' is the ensemble mean after DA. The analysis results indicate that the local filter (LESTKF) alters the model state mostly in the specified horizontal range (500 m).

The sensitivity to the observation error can be clearly seen in Fig. 4. Smaller observation errors lead to results closer to the 'observation', as expected. However, we remark that in general smaller observation errors do not always guarantee the analysis result is closer to the "truth" because 'observation' inevitably contains errors. In practice, observation errors must be set appropriately to achieve desired results. In summary, Figs. 3 and 4 indicate that the DA system works as intended.

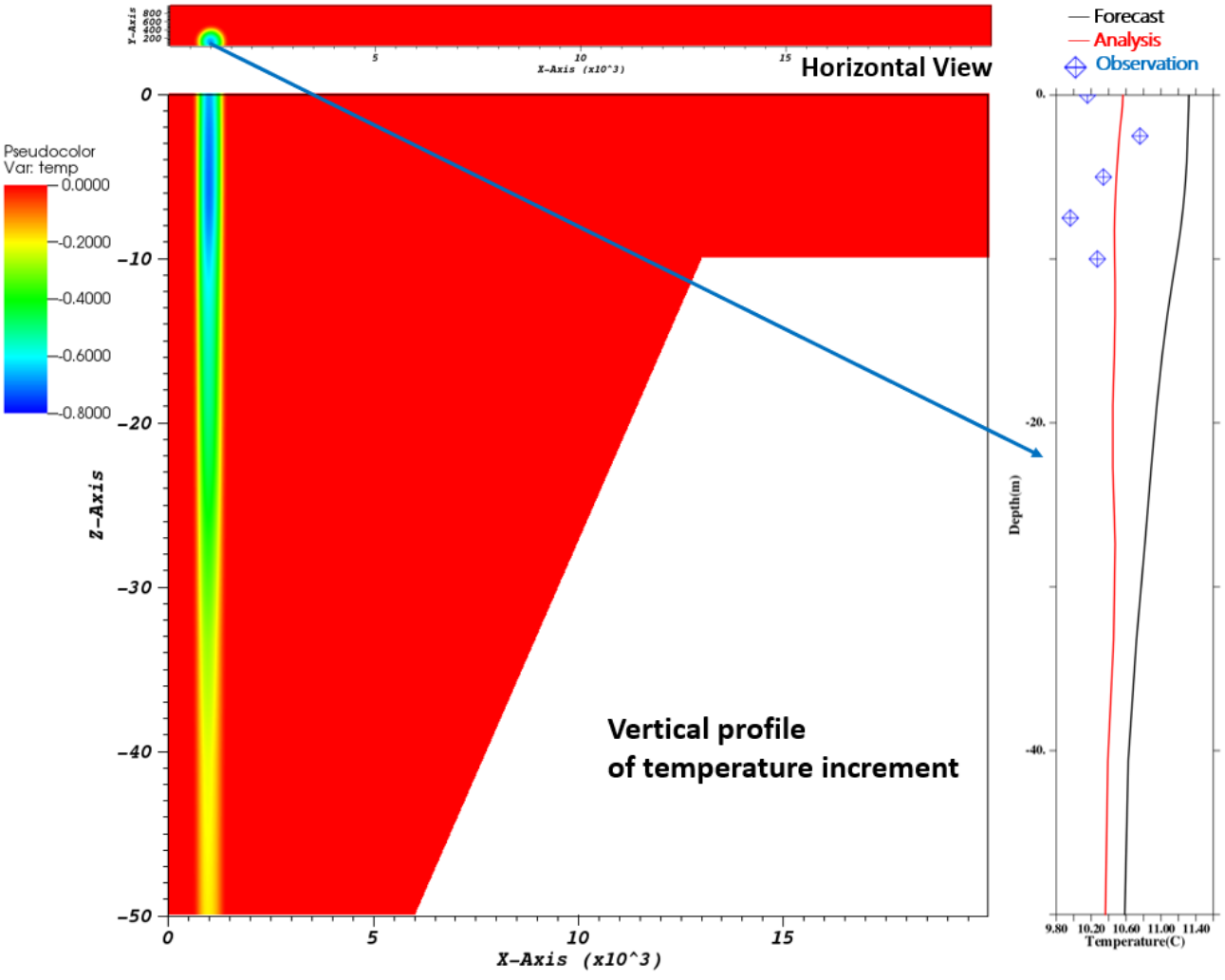

Fig 3. Lock-exchange test with 1 profile observation. The contour plots show the surface (top) and mid transect view (bottom) of the temperature difference before and after DA (i.e. Analysis-Forecast). The right-side plot shows the comparison of the temperature profile at the observation location.

https://doi.org/10.5194/egusphere-2022-114
Preprint. Discussion started: 20 May 2022

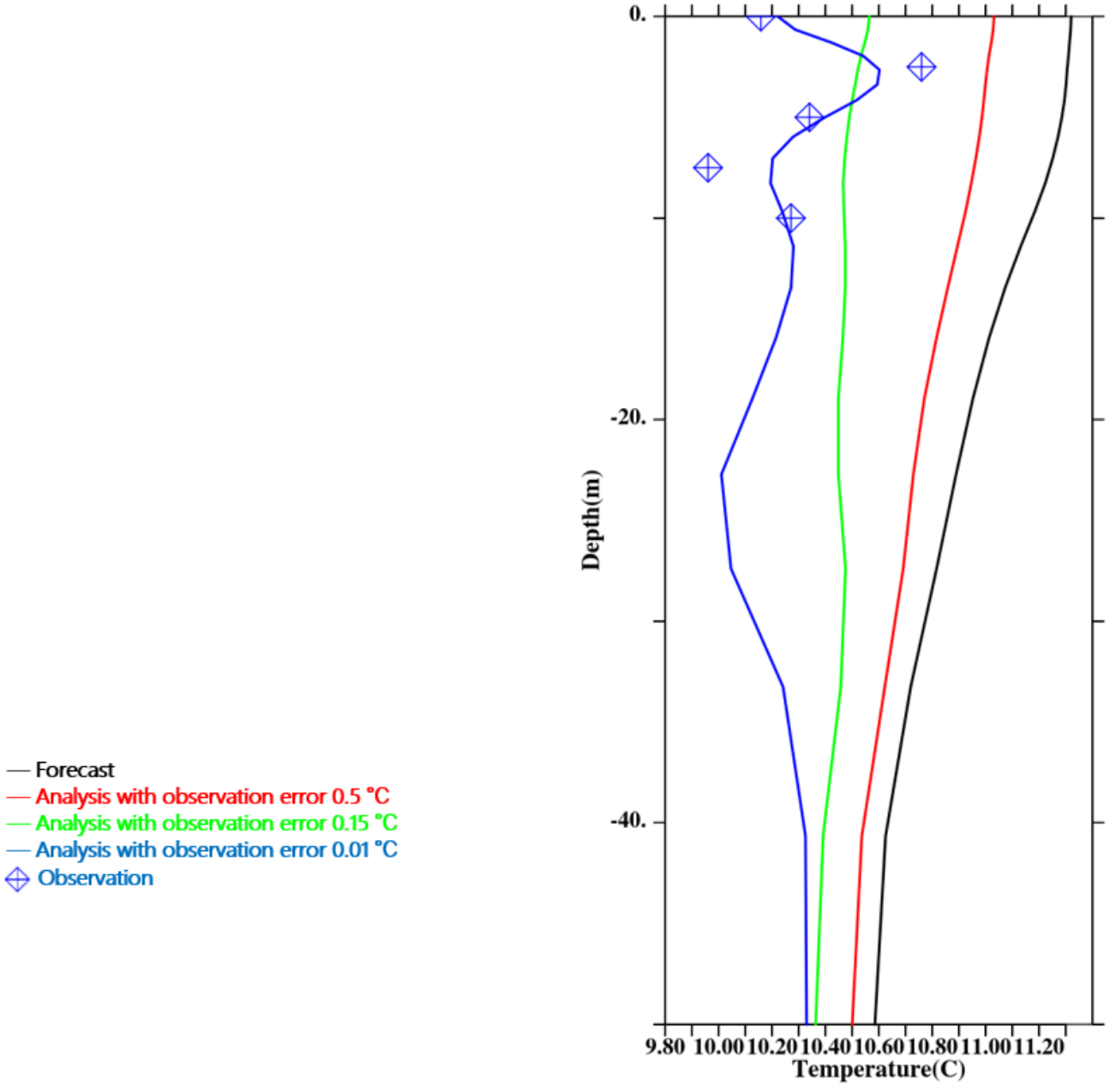

Fig 4. DA results with different observation errors.

3.1.4 Results from assimilating multiple profiles

In this setup, the temperature observation data points are extracted from 9 transects along the y-axis (y=100m, 200m, 300m …, 900m, every 100m) direction every 250 meters in the x direction as shown in the first panel of Fig. 5. In the vertical, observations are selected every 2.5 meters from surface to 40m depth. Here we test 4 different filters including 2 global

https://doi.org/10.5194/egusphere-2022-114
Preprint. Discussion started: 20 May 2022

(ETKF, ESTKF) and 2 local filters (LETKF, LESTKF). Using 8 ensemble members, all those filters can immediately correct the model bias at the first DA cycle (Fig. 5). Both global and local filters produce larger errors compared to the local filters (Fig. 5). Localization helps to limit observation effects within the analysis domain and the local filter analysis results at both ends of the domain (x< 6km and x> 13km) are shown to be closer to the truth. This difference can be understood from the fact that the global filter computes a global optimal ensemble combination with which the state estimate is incremented. This optimum is given by the relative errors of the model state and the observations. Since there are no observations at the ends of the domain, the assimilation results can deviate more there. In contrast the local filters decouple the computations of the optimums for distant locations from each other, so the increments on both ends of the domain are independent from each other. Our results confirm that the local filters tend to give better analysis results than the global filters. This effect of localization is well known for cases with abundant observations but also seems true for relatively sparse observations used here. Among the two local filters, LESTKF has certain advantages in that even if observation is locally missing for a long time, the inflation can still increase the ensemble spread for LETKF, but not the LESTKF (Nerger et al. 2012b), which may lead to over-amplified ensemble spread and thus worse analysis results with LETKF. Also, LESTKF is computationally slightly more efficient than LETKF. Thus, LESTKF is generally recommended for most applications.

https://doi.org/10.5194/egusphere-2022-114
Preprint. Discussion started: 20 May 2022

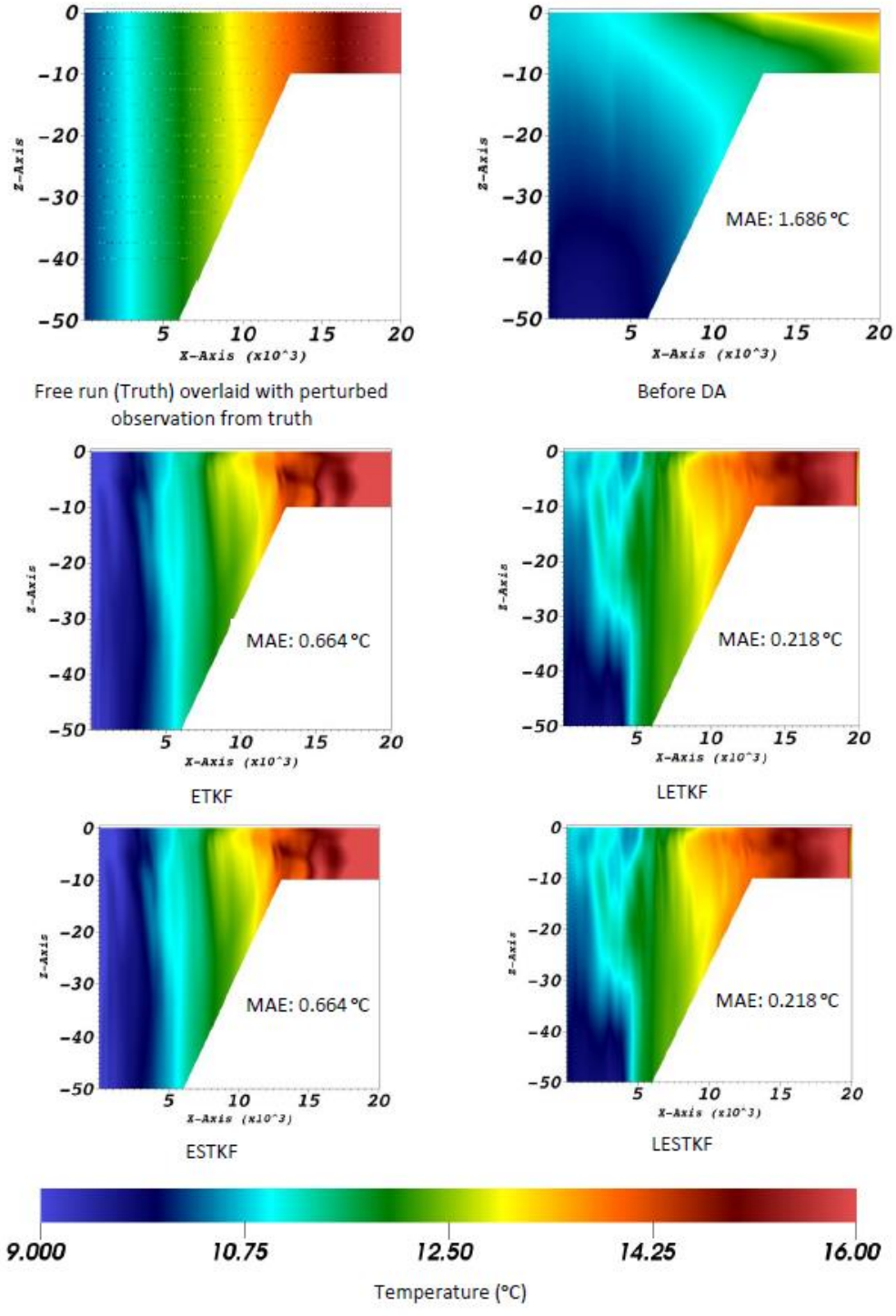

Fig 5. The vertical temperature transects after the 1st DA cycle using different filters. Initial model state is derived from 24-hour mean state with –1 ºC offset. Domain-wide averaged Mean Absolute Errors (MAEs) are shown for each case.

https://doi.org/10.5194/egusphere-2022-114
Preprint. Discussion started: 20 May 2022

3.2 Northwestern Pacific during typhoon events

3.2.1 DA setup

Northwestern Pacific is a very complex system with multiple current systems interacting with each other and with complex bathymetry and geometry (Johns et al. 2001; Jan et al. 2006). A major western boundary current, Kuroshio, plays a critical role in transporting warm and salty water from equatorial pacific to higher latitudes (Oey et al., 2013). Previously, we have successfully applied the 3D unstructured-grid model, SCHISM, to this challenging domain (Yu et al. 2017). In particular, we demonstrated that the combination of the flexible vertical gridding system with the flexible horizontal unstructured mesh is essential to capture many topographically driven processes.

The focus here is on circulations around Taiwan, which exhibit very complex features (Jan et al. 2006). The SCHISM setup can be found in Yu et al. (2017). HYCOM data is used for model initialization and open boundary conditions. The surface forcing is derived from ERA5 reanalysis surface field (Hersbach, H. et al., 2020). The simulation period starts from 18 June 2016 and ends on 2 August 2016 and covers Typhoon Nepartak, which made landfall near Taiwan around 7 July 2016. We selected this period to clearly show the improvement made by DA, as the forward model had larger errors.

Compared to the simple case shown in Section 3.1, this realistic case is much more challenging because (1) the observation is usually much sparser than model points and often distributed unequally in space; (2) different types of observation come with different uncertainties, which may adversely affect DA results. To illustrate the 2nd point, we compare SST observations at two coastal buoys with satellite SST observations (cf. Table 2). As shown in Fig. 6, the discrepancies between the two types of observations can be as large as 2.5℃. Therefore, we will not assimilate the buoy SST data. In general, conflicts between different types of observations represent a major hurdle for DA, in addition to poor data quality. Table 2 summarizes the observation types used in this section. ESA CCI-SST level-4 satellite analysis products are used as the main observation (Merchant et al., 2019). In addition, there are about 20-30 ARGO floats (Argo, 2021) in our domain depending on specific periods. Fig. 7 shows an example of ARGO distribution on 8 July 2016. In addition, AVISO sea level anomaly (SLA) from The Ssalto/Duacs altimeter products as produced and distributed by the Copernicus Marine and Environment Monitoring Service (CMEMS; https://www.copernicus.eu/en/copernicus-services/marine, last access: 10 May 2022) is combined with TPXO (Egbert and Erofeeva, 2002, https://www.tpxo.net/global, last access: 10 May 2022) tidal database to provide the observation for the total water elevation.

Two tests are shown here to illustrate the impact of different DA approaches (Table 3). Case A assimilates ESA CCI-SST with sparse ARGO data (about 20-30 vertical profiles each day) to correct the temperature. Case B additionally assimilates the sea level data from AVISO and TPXO. Validation for both cases will focus on SST and SLA around Taiwan. The observation error is provided by the data providers for the two satellite datasets. The observation errors of ARGO were set to 0.25℃ for temperature and 0.25 PSU for salinity to account for the small mismatches between ARGO and ESA CCI-SST.

The approach for model ensemble member generation follows that in the previous section, and we use 16 members in this test. We use LESTKF with a specified influence range of 0.5 degrees longitude and latitude.

https://doi.org/10.5194/egusphere-2022-114
Preprint. Discussion started: 20 May 2022

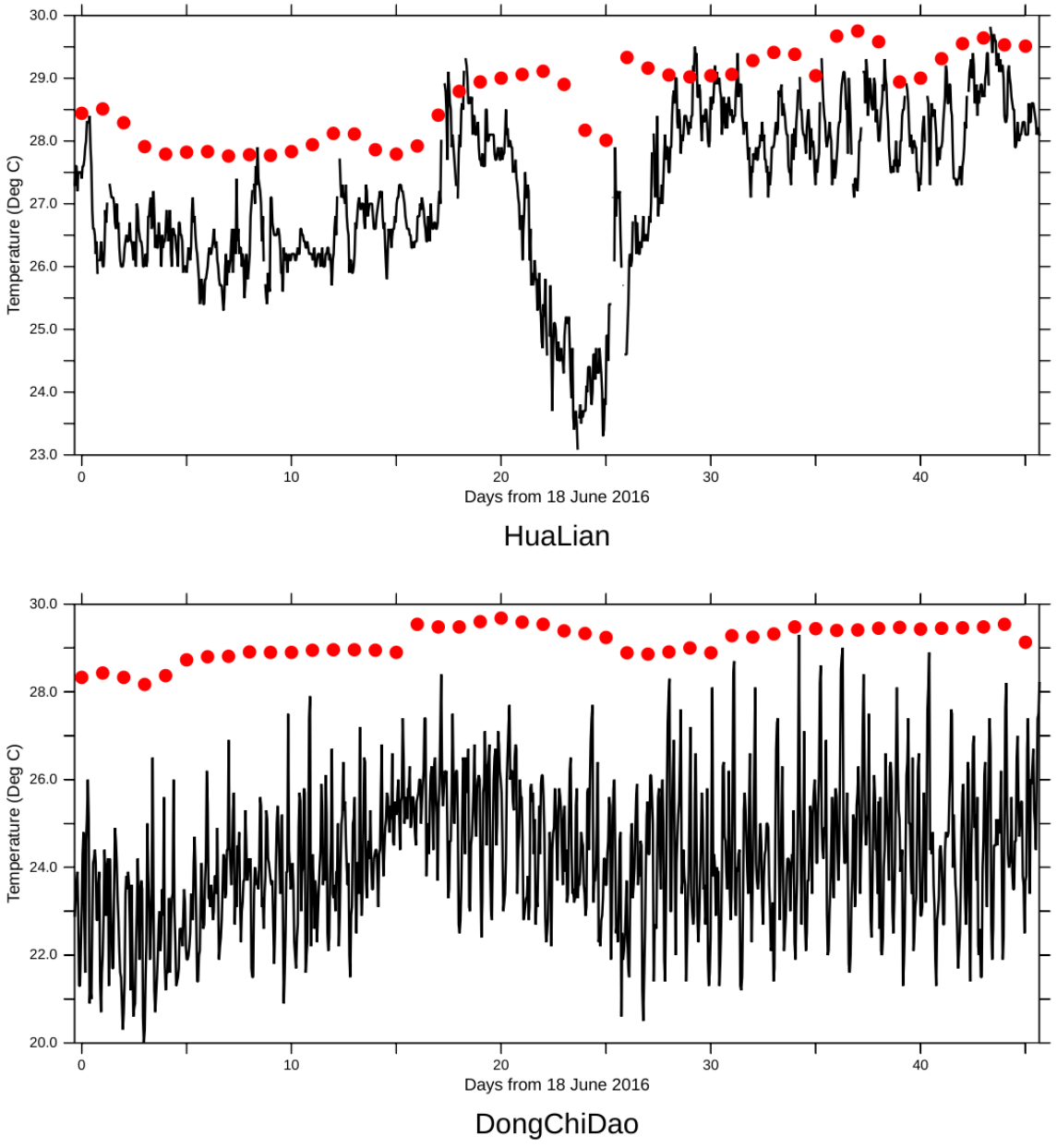

Fig. 6: Comparison of SST measured by in-situ buoys (blackline) and satellite (red dots) at two buoys near Taiwan. The buoy locations can be seen in Fig.7.

https://doi.org/10.5194/egusphere-2022-114
Preprint. Discussion started: 20 May 2022

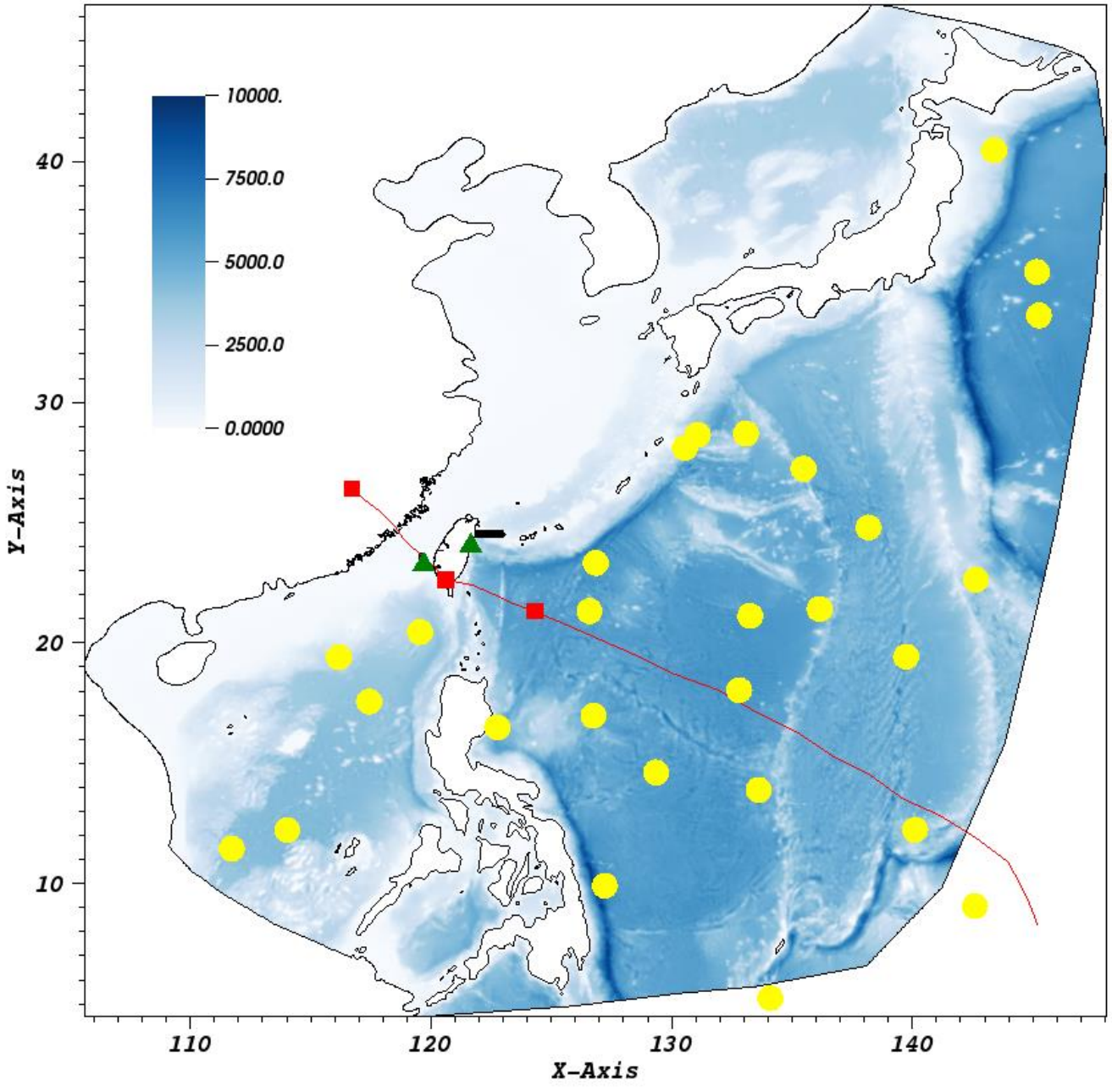

Fig. 7: Model domain for Northwestern Pacific with bathymetry, ARGO locations (yellow dots) on 8 July 2016, Central Weather Bureau buoy locations (green triangles), and Kuroshio transect location (black thick line), typhoon Nepartak track (red line), with the red squares representing the typhoon center locations corresponding to the times used in Figs. 9 & 11 (i.e., Day 19 (before), Day 20 (during), and Day 22 (after typhoon)).

https://doi.org/10.5194/egusphere-2022-114
Preprint. Discussion started: 20 May 2022

Table 2 Observations data types used

OBSERVATION TYPE	HORIZONTAL RESOLUTION	FREQUENCY	OBSERVATION ERROR
ESA LEVEL 4 CCI-SST	0.05 degree	Daily	Directly derived from error estimation (0.16 ~ 3.66℃)
ARGO	Sparse and random	Input with daily frequency	0.25 ℃ & 0.25 PSU
AVISO SLA + TPXO	0.25 degree	Daily	Directly derived from SLA error estimation (range: 0.01~0.14 m)

Table 3 Real case setup

CASE	OBSERVATION ASSIMILATED
A	ESA CCI-SST, ARGO
B	EST CCI-SST, ARGO, AVISO-SLA + TPXO

3.2.2 Discussion of assimilation results

Since both cases assimilate SST data, the simulated SST is improved immediately after the first DA cycle compared to the free run. This improvement is persistent throughout the entire 45-day simulation period, starting from relatively calm weather conditions to the brief severe typhoon and restoration period afterward. Fig. 8 shows that after 19 DA cycles, SST from both Cases A&B have similar improvement compared to the free run and remain close to the observation in the entire domain. Closer to Taiwan, Fig. 9 reveals some small differences between the two cases: assimilation of the SLA observation in Case B made a small difference in the deeper region, especially along the Kuroshio near Okinawa (in the region 124° E~128° E, 20° N~24° N). This difference persisted throughout the typhoon period (Fig. 9). Other than that, both cases captured well the Kuroshio induced upwelling near northeast corner of Taiwan. Overall, the two cases have similar SST skills (Fig. 10); however, Case A shows a slightly better score in the latter half of the simulation (Fig. 10), probably due to some minor compensating effects from assimilating the AVISO. The two DA results show significant improvement from the free run (Fig. 10).

https://doi.org/10.5194/egusphere-2022-114
Preprint. Discussion started: 20 May 2022

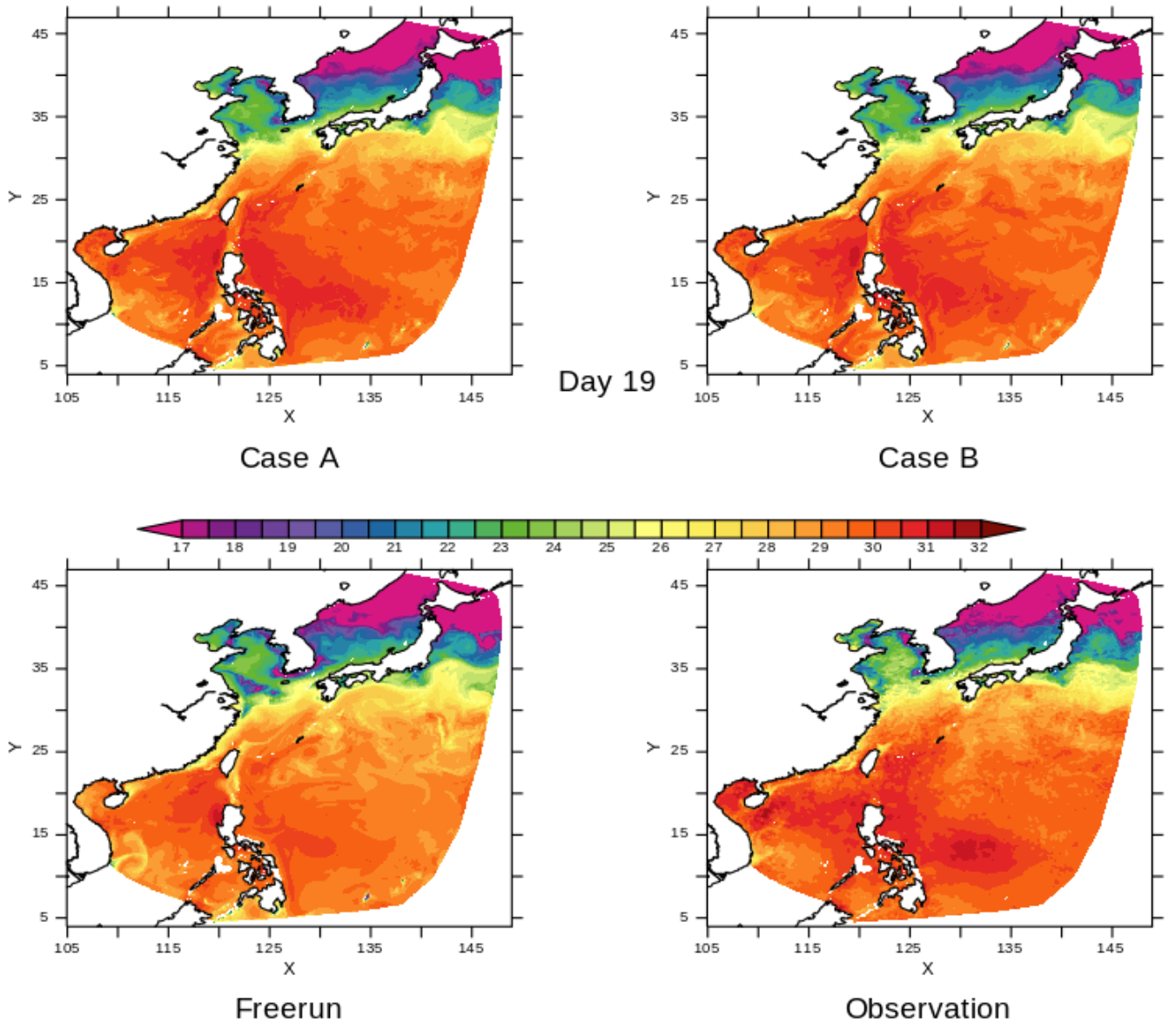

Fig. 8: SST comparison on Day 19 (7 July 2016, before typhoon).

https://doi.org/10.5194/egusphere-2022-114
Preprint. Discussion started: 20 May 2022

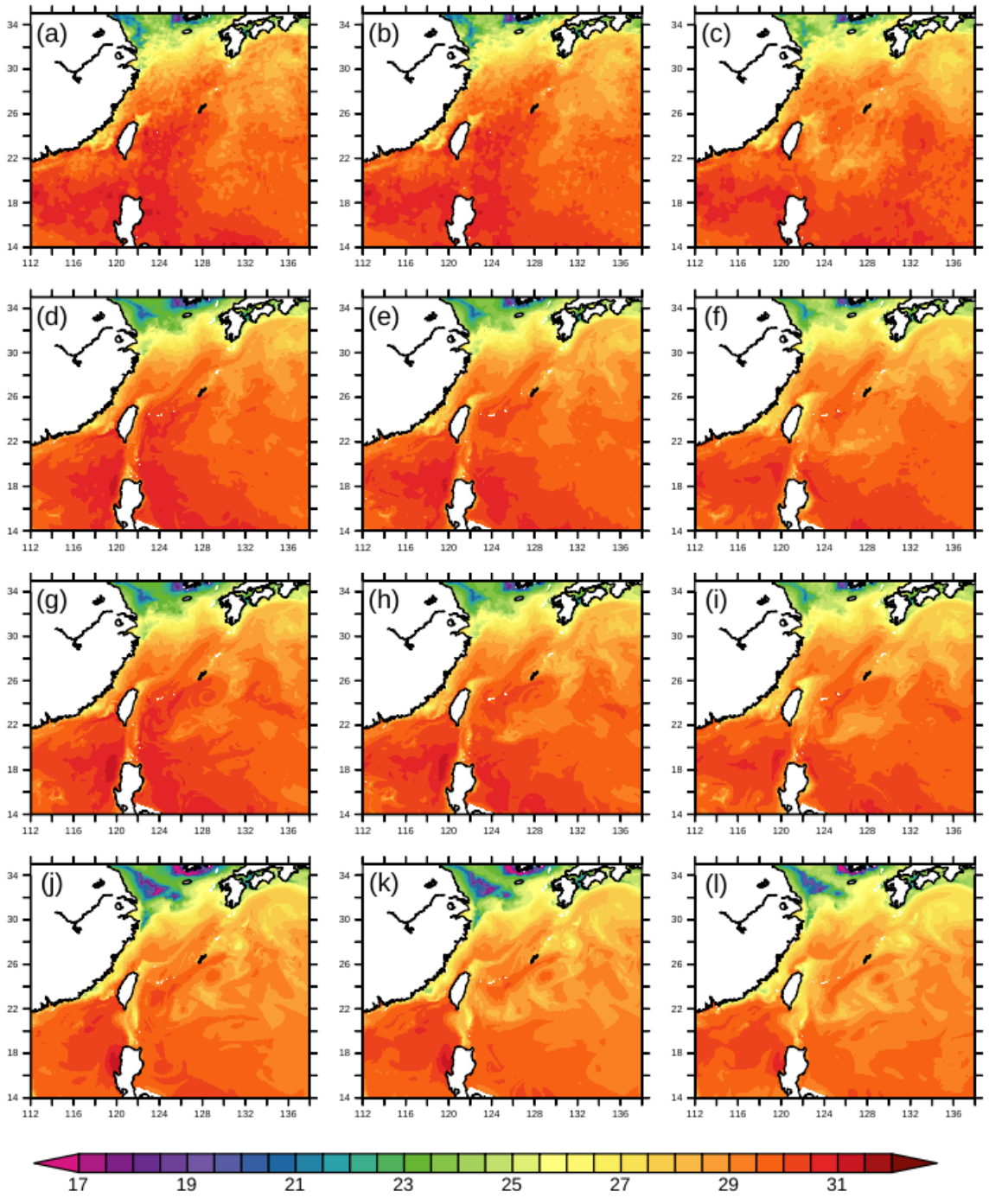

Fig. 9: SST comparison around Taiwan. Panels (a-c) represent observation on Day 19 (before typhoon, 7 July 2016), 20 (during typhoon, 8 July 2016),22 (after typhoon, 10 July 2016). The typhoon center locations are shown in Fig. 7. Panels (d-f) represent Case A; (g-i) represent Case B; (j-l) represent the Free run.

https://doi.org/10.5194/egusphere-2022-114
Preprint. Discussion started: 20 May 2022

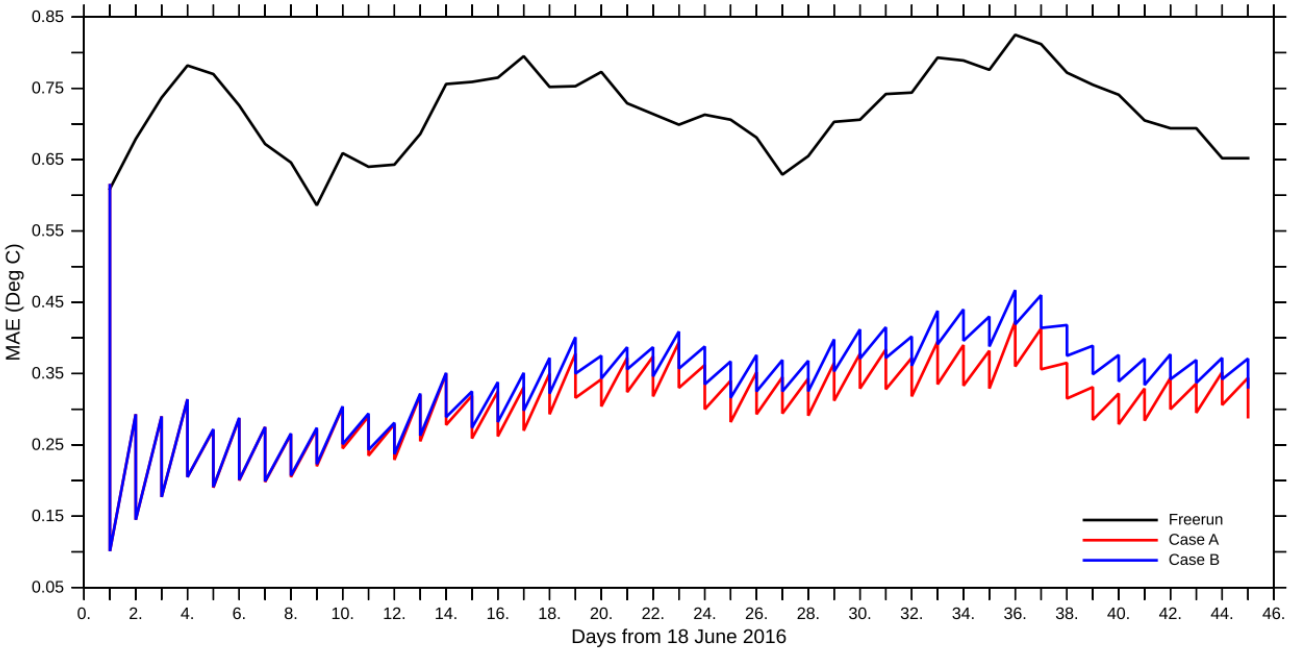

Fig 10: Comparison of averaged MAE of SST in the Taiwan region (shown in Fig. 9). The vertical jumps in cases A and B represent the DA effects at each analysis step.

For the SLA, both cases also deliver similar results. Case B, which explicitly assimilates AVISO, shows a slightly better performance and captures eddy location more accurately (Fig. 11). On the other hand, performance of Case A is also largely satisfactory (Fig. 11). Note that for Case A, the SLA data can be considered as an independent validation dataset to evaluate how assimilating the temperature observation may improve other variables. Results from Cases A&B show improved skill from the free run in all stages of the typhoon (Fig. 11). As expected, the SLA in Case B has the best skill among all cases, especially after 10 days of DA (Fig. 12).

https://doi.org/10.5194/egusphere-2022-114
Preprint. Discussion started: 20 May 2022

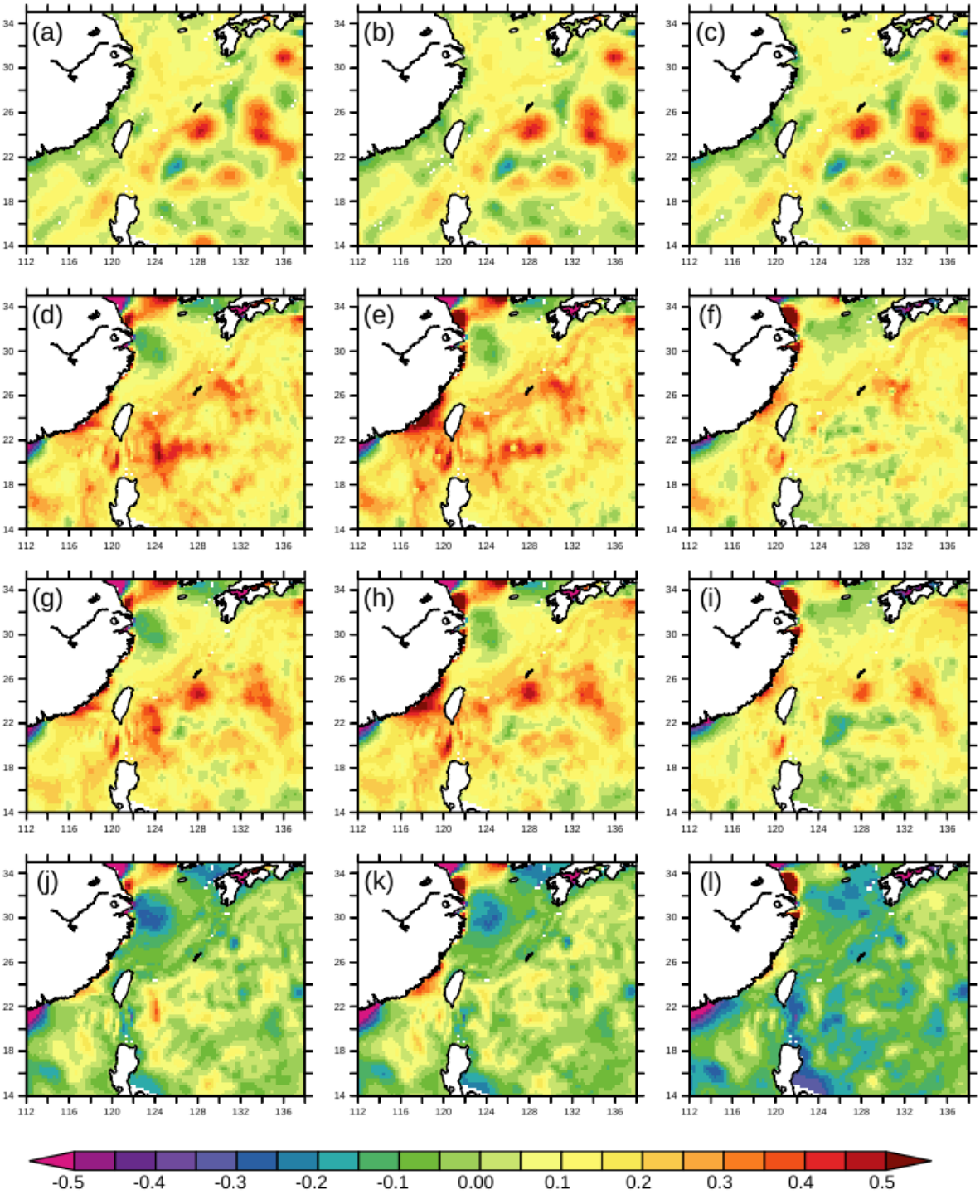

Fig 11: SLA comparison around Taiwan. The three columns correspond to before, during and after the typhoon. (a-c): observation; (d-f): Case A; (g-i): Case B; (j-l): Free run.

https://doi.org/10.5194/egusphere-2022-114
Preprint. Discussion started: 20 May 2022

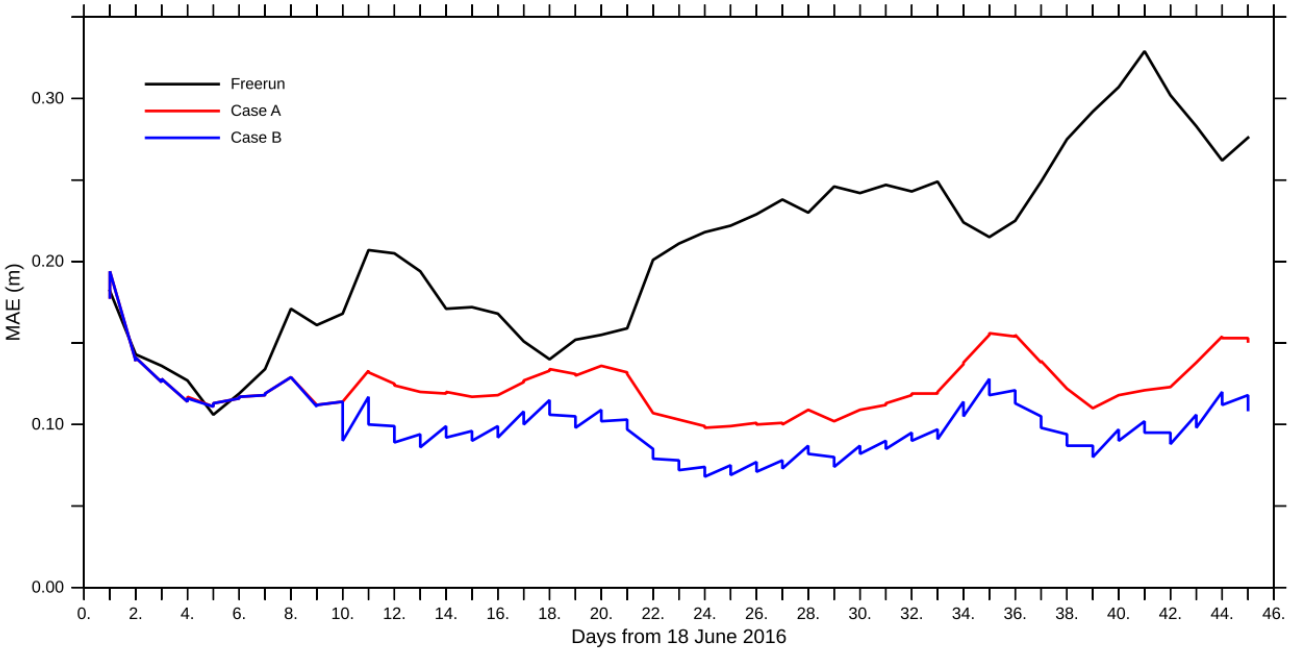

Fig. 12: Comparison of averaged MAE for SLA in the subdomain region around Taiwan; the jumps in Case B correspond to before/after DA step.

The DA has also improved the model skill for Kuroshio transport. As shown in Fig. 13, both Cases A&B manage to increase the Kuroshio velocity over the top 700 m, which in turn has enhanced the Kuroshio transport near Taiwan (Fig. 14), nudging it closer to the climatological value of 23 Sv (Johns et al. 2001). The improved model skill can be attributed to the increased SST after DA, which has enhanced the density gradients (Fig. 9).

https://doi.org/10.5194/egusphere-2022-114
Preprint. Discussion started: 20 May 2022

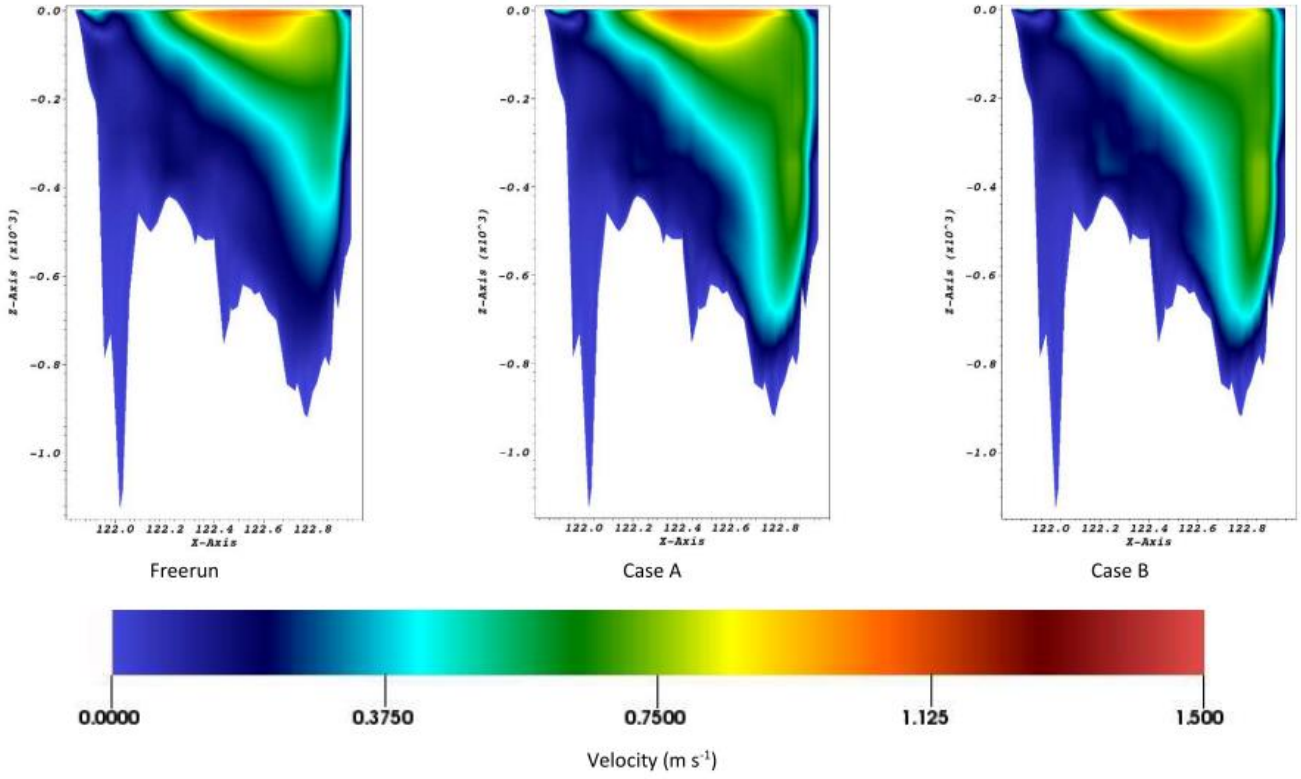

Fig. 13: Normal velocity along a transect along the Kuroshio path at 24.5° N (see Fig. 7 for its location) on Day 19 from different cases.

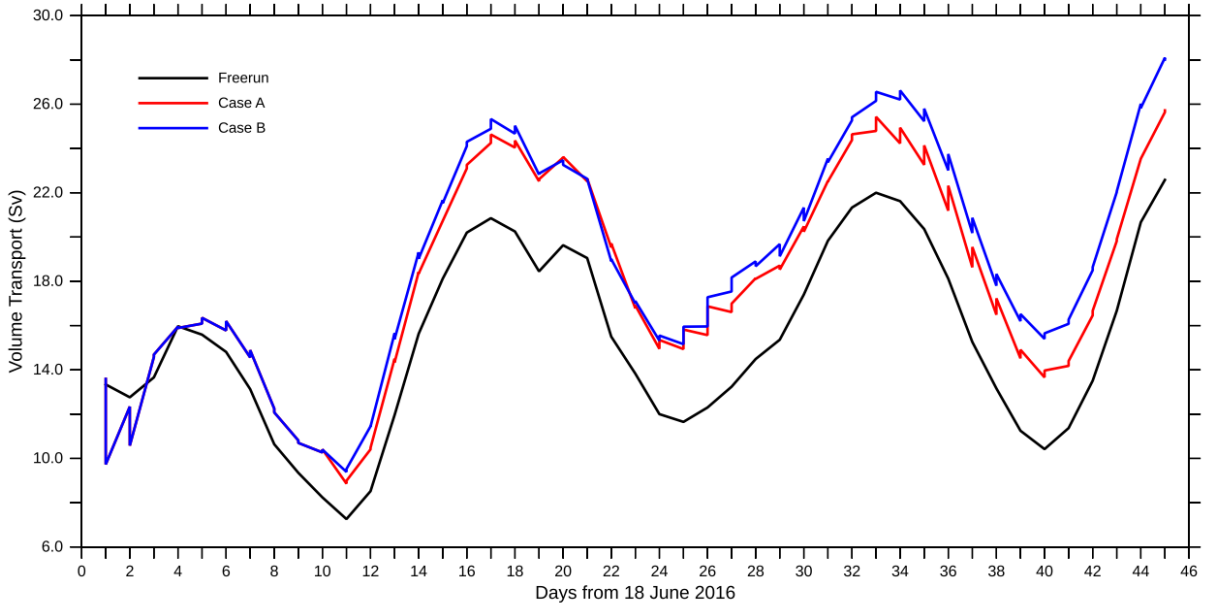

Fig. 14: Time series of the total Kuroshio transport from different cases along a transect (see Fig. 7 for its location).

https://doi.org/10.5194/egusphere-2022-114
Preprint. Discussion started: 20 May 2022

3.3 Code efficiency

PDAF allows maximum flexibility in setting up the DA experiments in either 'flexible' or 'full parallel' modes, based on availability of computational resources. If sufficient resources are available, the 'fully parallel' mode is the most efficient option. Under this mode, all ensemble members are executed concurrently (Fig. 2). In reality, however, computational resources are usually limited, and the flexible mode is often the practical option. Under this mode, the user can specify how many members are executed concurrently in batches, and memory sharing and rewinding of clock are necessary between different cohorts (Fig. 2). Obviously, one way to maximize efficiency is to minimize the number of cohorts.

In this section we test the performance using different numbers of ensemble members under the fully parallel and flexible modes using the same test shown in Section 3.2. The simulations were conducted on Extreme Science and Engineering Discovery Environment (XSEDE)'s Frontera (https://frontera-portal.tacc.utexas.edu/user-guide/, last access: 10 May 2022). Table 4 summarizes the simulation times from different tests. The results indicate that there is no significant time difference between different numbers of ensemble members under the fully parallel mode, which demonstrates the excellent scaling and minimal overhead induced by PDAF and ESMF. Both ensemble forecast and assimilation analysis steps took similar times. Using the flexible mode, the ensemble DA time is very close to that in the corresponding run using the fully parallel mode after factoring in the number of cohorts (=4), and the time spent on the assimilation analysis step is about the same as that in the fully parallel mode. This again illustrates minimal overhead induced by PDAF and ESMF. The code can efficiently handle both modes and thus give users options to adjust the cost of ensemble simulation based on available resources. Compared to the free run, the ensemble forecast step using fully parallel mode only induces less than 4% overhead cost. Furthermore, the overhead is negligible under the flexible mode for the forecast step (after factoring in the number of cohorts). The cost for the analysis step is mainly dependent on the amount of observation data as each observation data point needs to be searched within the local analysis domain, which accounts for about 93% of the cost for the analysis step. In other words, the PDAF filter only accounts for ~7% of the cost in the analysis step, and the search algorithm is the bottleneck for DA analysis. This bottleneck will be further improved in the future using newer versions of PDAF. Overall, the DA tool adds a modest (~20%) overhead to the free run in the total time.

https://doi.org/10.5194/egusphere-2022-114
Preprint. Discussion started: 20 May 2022

Table 4. Wall-clock time (for 28 days of simulation) from using different numbers of ensemble members under fully parallel and flexible mode on Frontera (with 56 cores/node).

Ensemble members	Compute Nodes	Cores	Init (mins)	Ensemble forecast (mins)	DA analysis (mins)	Total (mins)
Freerun	10	560	0.5	306	0	306.5
8 (fully parallel)	80	4480	0.5	308.3	54.4	363.2
16 (fully parallel)	160	8960	0.5	312.5	55.7	368.7
24 (fully parallel)	240	13440	0.5	312.55	55.5	368.55
8 (flexible mode, 2 members in each cohort)	20	1120	0.5	1219.98	56.4	1276.88

330

4. Conclusions

We have developed a new data assimilative (DA) system by combining two parallel frameworks: a parallel DA framework (PDAF) and a flexible model coupling framework (ESMF). The new DA system is built on the ESMF at the top level that drives the PDAF and any combination of earth system modeling (ESM) components. In addition, the new DA system
335 supports the two operating modes of PDAF: full parallel (when sufficient resources are available) or flexible (when resources are insufficient). Therefore, the new system allows maximum flexibility and easy implementation of data assimilation for fully coupled ESM applications. The new system was successfully applied to a realistic case of Kuroshio simulation around Taiwan using remote sensed and in-situ observation, and it significantly improved the model skill for temperature, velocity and surface elevation before, during and after typhoon events. Future work will extend the current
340 system to include other types of observations (e.g., velocity measurement from Coastal Ocean Dynamics Applications Radar (CODAR), Acoustic Doppler Current Profiler (ADCP)) and coupled ESM applications.

https://doi.org/10.5194/egusphere-2022-114
Preprint. Discussion started: 20 May 2022

Author contribution

All authors contributed to writing of the paper. HCY, YJZ, CL and LN developed the code with the support of other co-authors. LN provided valuable guidance on using PDAF. CHC and CTT designed the tests and assisted in the interpretation of results. HCY conducted the simulations with support from YJZ. JCSY and TYC provided the observational data.

Competing interests

The authors declare that they have no conflict of interest.

Acknowledgement

This work is funded by the Central Weather Bureau (Taiwan) under grant 110A034-1, and by Bundesministerium für Bildung und Forschung (BMBF) grant Multiple Stressors on North Sea Life (MuSSeL, 03F0862A). Simulations used in this paper were conducted using the following computational facilities:

(1) William & Mary Research Computing for providing computational resources and/or technical support (URL: https://www.wm.edu/it/rc, last access: 10 May 2022)

(2) the Extreme Science and Engineering Discovery Environment (XSEDE; Grant EAR21010), which is supported by National Science Foundation grant number OCI-1053575;

(3) Texas Advanced Computing Center (TACC), The University of Texas at Austin.

Code availability

The software package used in this study is archived at Zenodo (https://zenodo.org/record/6535437). The PDAF code (version 1.16 was used here) is also archived at Zenodo (https://doi.org/10.5281/zenodo.6535821), a full code documentation and a usage tutorial, is available at http://pdaf.awi.de (last access: 10 May 2022). This package requires ESMF (version 8.1.0 was used here) which are freely available at https://earthsystemmodeling.org/download/ (last access: 10 May 2022). Source code of SCHISM can be accessed at https://zenodo.org/record/6537527.

Data availability

The ARGO data was downloaded from ARGO project (http://doi.org/10.17882/42182). Detail description can be found at https://argo.ucsd.edu/data/acknowledging-argo/(last access: 10 May 2022). ESA CCI-SST data can be downloaded from Centre for Environmental Data Analysis (CEDA) archive (Merchant et al., 2019, http://dx.doi.org/10.5285/62c0f97b1eac4e0197a674870afe1ee6). AVISO SLA data can be downloaded from CEMES

https://doi.org/10.5194/egusphere-2022-114
Preprint. Discussion started: 20 May 2022

archive (https://doi.org/10.48670/moi-00148). TPXO is maintained by CEOAS, Oregon State University, USA (https://www.tpxo.net/global, last access: 10 May 2022) and freely available for academic research usage.

References

Argo: Argo float data and metadata from Global Data Assembly Centre (Argo GDAC). SEANOE. https://doi.org/10.17882/42182, 2021.

Bannister, R.: A review of operational methods of variational and ensemble-variational data assimilation, Q. J. R. Meteorol. Soc. 143: 607 – 633. https://doi.org/10.1002/qj.2982, 2017.

Brune, S., Nerger, L., Baehr, J.: Assimilation of oceanic observations in a global coupled Earth system model with the SEIK filter Ocean Modelling. Ocean Modelling, 96, 254-264 doi:10.1016/j.ocemod.2015.09.011, 2015.

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

Evensen, G.: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. Journal of Geophysical Research 99 (C5), 10143–10162, 1994.

Fournier, A., Nerger, L., Aubert, J.: An ensemble Kalman filter for the time-dependent analysis of the geomagnetic field. Geochemistry, Geophysics, Geosystems, 14, 4035-4043 doi:10.1002/ggge.20252, 2013.

Goodliff, M., Bruening, T., Schwichtenberg, F., Li, X., Lindenthal, A., Lorkowski, I., Nerger, L.: Temperature assimilation into a coastal ocean-biogeochemical model: Assessment of weakly and strongly-coupled data assimilation, Oce. Dyn., 69, 1217-1237, doi:10.1007/s10236-019-01299-7, 2019.

Gropp, W., Lusk, E., Skjellum, A.: Using MPI—Portable Parallel Programming with the Message-Passing Interface. The MIT Press, Cambridge, 1994.

Hersbach, H, Bell, B., Berrisford, P., Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, Adrian Simmons, Cornel Soci, Saleh Abdalla, Xavier Abellan, Gianpaolo Balsamo, Peter Bechtold, Gionata Biavati, Jean Bidlot, Massimo Bonavita, Giovanna De Chiara, Per Dahlgren, Dick Dee, Michail Diamantakis, Rossana Dragani, Johannes Flemming, Richard Forbes, Manuel Fuentes, Alan Geer, Leo Haimberger, Sean Healy, Robin J. Hogan, Elías Hólm, Marta Janisková, Sarah Keeley, Patrick Laloyaux, Philippe Lopez, Cristina Lupu, Gabor Radnoti, Patricia de Rosnay, Iryna Rozum, Freja Vamborg, Sebastien Villaume, Jean-Noël Thépaut: The ERA5 global reanalysis. Q J R Meteorol Soc. 146: 1999– 2049. https://doi.org/10.1002/qj.3803, 2020.

Egbert, Gary D., and Svetlana Y. Erofeeva.: Efficient inverse modeling of barotropic ocean tides, Journal of Atmospheric and Oceanic Technology 19.2, 183-204, 2002.

Hunt, B. R., E. J. Kostelich, and I. Szunyogh: Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter. Physica D., 230, 112–126, 2007.

https://doi.org/10.5194/egusphere-2022-114
Preprint. Discussion started: 20 May 2022

Jan, S., Sheu, D.D., Kuo, H.-M.: Water mass and throughflow transport variability in the Taiwan Strait. J. Geophys. Res. 111, C12012. http://dx.doi.org/10.1029/2006JC003656, 2006.

Johns, W.E., Lee, T.N., Zhang, D., Zantopp, R.: The Kuroshio East of Taiwan: Moored transport observations from the WOCE PCM-1 array. J. Phys. Oceanogr. 31,1031–1053, 2001.

Kalnay, E., Li,H., Miyoshi,T., Yang, S., Ballabrera-Poy, J.: 4-D-Var or ensemble Kalman filter? Tellus, 59a, 758-773, doi: 10.1111/j.1600-0870.2007.00261.x, 2007.

Kurtz, W., G. He, S. J. Kollet, R. M. Maxwell, H. Vereecken, H.-J. Hendrics Franssen: 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.

Lemmen, C., Hofmeister, R., Klingbeil, K., Nasermoaddeli, M. H., Kerimoglu, O., Burchard, H., Kösters, F., and Wirtz, K. W.: Modular System for Shelves and Coasts (MOSSCO v1.0) – a flexible and multi-component framework for coupled coastal ocean ecosystem modelling, Geosci. Model Dev., 11, 915–935, https://doi.org/10.5194/gmd-11-915-2018, 2018.

Merchant, C.J., Embury, O., Bulgin, C.E., Block T., Corlett, G.K., Fiedler, E., Good, S.A., Mittaz, J., Rayner, N.A., Berry, D., Eastwood, S., Taylor, M., Tsushima, Y., Waterfall, A., Wilson, R., Donlon, C.: Satellite-based time-series of sea-surface temperature since 1981 for climate applications, Scientific Data 6:223, http://doi.org/10.1038/s41597-019-0236-x, 2019.

Mu, L., Liang, X., Yang, Q., Liu, J., Zheng, F.: Arctic Ice Ocean Prediction System: evaluating sea-ice forecasts during Xuelong's first trans-Arctic Passage in summer 2017. Journal of Glaciology 1–9, doi:10.1017/jog.2019.55, 2019.

Nerger, L., Hiller, W., Schröter, J.: PDAF - The Parallel Data Assimilation Framework: Experiences with Kalman Filtering, Use of high performance computing in meteorology : proceedings of the Eleventh ECMWF Workshop on the Use of High Performance Computing in Meteorology, Reading, UK, 25 - 29 October 2004 / Eds.: Walter Zwieflhofer; George Mozdzynski, Singapore: World Scientific, 63-83. doi:10.1142/9789812701831_0006, 2005.

Nerger, L., Danilov, S., Hiller, W., Schröter, J.: Using sea-level data to constrain a finite-element primitive-equation ocean model with a local SEIK filter, Ocean Dynamics, 56, 634-649, doi:10.1007/s10236-006-0083-0., 2006.

Nerger, L., Danilov, S., Kivman, G., Hiller, W., Schröter, J.: Data assimilation with the Ensemble Kalman Filter and the SEIK filter applied to a finite element model of the North Atlantic, Journal of Marine Systems, 65(1/4), 288-298., doi:10.1016/j.jmarsys.2005.06.009, 2007.

Nerger, L., Janjić, T., Schröter, J., Hiller, W.: A unification of ensemble square root Kalman filters. Monthly Weather Review, 140, 2335-2345, 2012a.

Nerger, L., Janjić, T., Schröter, J., Hiller, W.: A regulated localization scheme for ensemble-based Kalman filters. Quarterly Journal of the Royal Meteorological Society, 138, 802-812, 2012b.

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

https://doi.org/10.5194/egusphere-2022-114
Preprint. Discussion started: 20 May 2022

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

Oey, L., Chang, Y., Lin, Y., Chang, M., Xu, F., Lu, H.: ATOP – Advanced Taiwan Ocean Prediction System based on the mpiPOM. Part 1: model descriptions, analyses and results. Terr. Atmos. Ocean. Sci. 24 (1), 137–158. http://dx.doi.org/10.3319/TAO.2012.09.12.01(Oc), 2013.

Ott, L.E., J.T. Bacmeister, S. Pawson, K.E. Pickering, G. Stenchikov, M.J. Suarez, H. Huntreiser, M. Loewenstein, J. Lopez, I. Xueref-Remy: An Analysis of Convective Transport and Parameter Sensitivity in a Single Column Version of the Goddard Earth Observing System, Version 5, General Circulation Model, J. Atmos. Sci., 66, 627-646. doi: 10.1175/2008JAS2694.1., 2009.

Pardini, F., Corradini, S., Costa, A., Esposti Ongaro, T., Merucci, L., Neri, A., Stelitano, D., de' Michieli Vitturi, M.: Ensemble-based data assimilation of volcanic ash clouds from satellite observations: Application to the 24 December 2018 Mt. Etna explosive eruption. Atmosphere, 11 359 doi:10.3390/atmos11040359, 2020.

Pham, D.T., Verron, J., Gourdeau, L.: Singular evolutive Kalman filters for data assimilation in oceanography. Comptes Rendus de l'Acade´mie die Science Paris, Series II 326 (4), 255–260, 1998a.

Pham, D.T., Verron, J., Roubaud, M.C.: A singular evolutive extended Kalman filter for data assimilation in oceanography. Journal of Marine Systems 16, 323–340, 1998b.

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

Pradhan, H.K., Voelker, C., Losa, S.N., Bracher, A., Nerger, L.: Assimilation of global total chlorophyll OC-CCI data and its impact on individual phytoplankton fields. J. Geophys. Res. Oceans, 124, 470-490, doi:10.1029/2018JC014329, 2019.

Pradhan, H.K., Voelker, C., Losa, S.N., Bracher, A., Nerger, L.: Global assimilation of ocean-color data of phytoplankton functional types: Impact of different datasets. J. Geophys. Res. Oceans, 125, e2019JC015586 doi:10.1029/2019JC015586, 2020.

Schachtschneider, R., J. Saynisch-Wagner, V. Klemann, M. Bagge, M. Thomas: An approach for constraining mantle viscosities through assimilation of palaeo sea level data into a glacial isostatic adjustment model. Nonlinear Processes in Geophysics 29, 53-75 doi:10.5194/npg-29-53-2022, 2022.

Tödter, J., and B. Ahrens: A second-order exact ensemble square root filter for nonlinear data assimilation. Monthly Weather Review, 143, 1347–1367, 2015.

Valcke, S.; Piacentini, A.; Jonville, G.: Benchmarking Regridding Libraries Used in Earth System Modelling. Math. Comput. Appl., 27, 31. https://doi.org/10.3390/mca27020031, 2022.

Vetra-Carvalho, S., van Leeuwen, P. J., Nerger, L., Barth, A., Altaf, M. U., Brasseur, P., Kirchgessner, P., Beckers, J.-M.: State-of-the-art stochastic data assimilation methods for high-dimensional non-Gaussian problems. Tellus A, 70:1, 1445364, doi:10.1080/16000870.2018.1445364, 2018.

https://doi.org/10.5194/egusphere-2022-114
Preprint. Discussion started: 20 May 2022

Yang, Q., Losa, S. N., Losch, M., Tian-Kunze, X., Nerger, L., Liu, J., Kaleschke, L., Zhang, Z.: Assimilating SMOS sea ice thickness into a coupled ice-ocean model using a local SEIK filter. Journal of Geophysical Research-Oceans, 119, 6680-6692, doi:10.1002/2014JC009963, 2014.

Yu, H-C., Zhang, Y., Yu, J.C.S., Terng, C., Sun, W., Ye, F., Wang, H.V., Wang, Z., and Huang, H.: Simulating multi-scale oceanic processes around Taiwan on unstructured grids, Ocean Modelling, 112, 72-93. http://dx.doi.org/10.1016/j.ocemod.2017.09.007, 2017.

Zhang, Y., Ateljevich, E., Yu, H-C., Wu, C-H., and Yu, J.C.S.: A new vertical coordinate system for a 3D unstructured-grid model, Ocean Modelling, 85, 16-31, 2015.

Zhang, Y., Ye, F., Stanev, E.V., Grashorn, S.: Seamless cross-scale modeling with SCHISM, Ocean Modelling, 102, 64-81. doi:10.1016/j.ocemod.2016.05.002, 2016.