scientific reports

Check for updates

OPEN

Assimilation of ground-based GNSS data using a local ensemble Kalman filter

Changliang Shao[1]✉ & Lars Nerger[2]

Tropical cyclones become increasingly nonlinear and dynamically unstable in high-resolution models. The initial conditions are typically sub-optimal, leaving scope to improve the accuracy of forecasts with improved data assimilation. Simultaneously, the lack of real ground-based GNSS observations over the ocean poses significant challenges when evaluating the assimilation results in oceanic regions. In this study, an Observation System Simulation Experiment is carried out based on a tropical cyclone case. Assimilation experiments using the WRF-PDAF framework are conducted. Conventional and GNSS observation operators are implemented. A diverse array of synthetic observations, encompassing temperature (T), wind components (U and V), precipitable water (PW), and zenith total delay (ZTD), are assimilated utilizing the Local Error-Subspace Transform Kalman filter (LESTKF). The findings highlight the improvement in forecast accuracy achieved through the assimilation process over the ocean. Multiple observation types further improve the forecast accuracy. The study underscores the crucial role of GNSS data assimilation techniques. The assimilation of GNSS data presents potential for advancing weather forecasting capabilities. Thus, the construction of ground-based GNSS observation stations over the ocean is promising.

Keywords Data Assimilation, LESTKF, Ground-based GNSS, Observation operator, Precipitable water, zenith total delay

Data assimilation (DA) plays a crucial role in improving the accuracy and reliability of numerical weather prediction (NWP) models. DA helps to bridge the gap between model simulations and real-world observations. It enhances the accuracy, skill, and reliability of the atmosphere simulations, providing valuable information for a range of applications, including weather forecasting, climate studies, and environmental assessments[1,2]. Global Navigation Satellite System (GNSS) data, such as those provided by the Global Positioning System (GPS), Galileo navigation satellite system (Galileo), global navigation satellite system (GLONASS), and BeiDou navigation satellite system (BDS), have gained significant attention in recent years due to their potential for improving atmospheric models and weather forecasting. Assimilating GNSS data into atmospheric models using techniques like Ensemble Kalman Filtering (EnKF) has shown significant impact on improving the accuracy and performance of the models[3] in different applications as outlined below.

Improved Initial Conditions: Assimilating GNSS data helps in refining the initial conditions of atmospheric models by incorporating real-time and high-resolution (e.g., 5-min series in the Nevada Geodetic Laboratory (NGL)[4] product (http://geodesy.unr.edu, last accessed: 27 July 2024) information about atmospheric parameters such as water vapor content. This leads to a better representation of the current state of the atmosphere and reduces the uncertainties associated with the initial conditions[5]. Positive impacts on weather forecasting, particularly for short-term (up to 6 h)[6] and short-range (48 h)[7] forecasts have been shown. The assimilation helps in capturing mesoscale weather phenomena such as convective systems, thunderstorms, and localized rainfall patterns[8]. It contributes to the better representation of atmospheric processes and improves the skill of weather forecasts, especially in regions where traditional observations are sparse or limited.

Enhanced Moisture Analysis: GNSS data assimilation plays a crucial role in improving moisture analysis in atmospheric models. It provides high-temporal and spatial resolution observations of precipitable water vapor and zenith total delay, which are vital for understanding the moisture distribution in the atmosphere. Assimilating GNSS data leads to a more accurate representation of moisture fields, enabling improved forecasts of precipitation and humidity patterns[9].

[1]CMA Research Centre On Meteorological Observation Engineering Technology, CMA Meteorological Observation Centre, Beijing, China. [2]Alfred-Wegener-Institut, Helmholtz-Zentrum Für Polar- Und Meeresforschung (AWI), Bremerhaven, Germany. ✉email: shaocl@cma.gov.cn

Vertical Profiling: GNSS data assimilation enables improved vertical profiling of atmospheric parameters, which is crucial for understanding the vertical structure of the atmosphere. By assimilating GNSS data, the vertical distribution of water vapor and other variables can be accurately estimated, aiding in the analysis of atmospheric stability, moisture transport, and cloud formation processes[10].

Real-time Assimilation: One of the key advantages of GNSS data is its availability in real-time. Assimilating GNSS data in real-time allows for timely updates of atmospheric models, leading to improved nowcasting and short-term/range forecasts. Data of zenith total delay (ZTD) and precipitable water (PW) are available within 1 h from the moment the original satellite signal is received by the ground-based station. This latency period encompasses both data processing and transmission times. Real-time assimilation enables the models to capture rapidly changing atmospheric conditions, providing valuable information for severe weather events and rapid weather developments[11].

Some of the most important challenges for the assimilation of GNSS data include the required greater sophistication of forward models to allow using the indirect observations PW and ZTD, the need to analyze a range of hydrometeors, the need to account for the flow-dependent multivariate "balance" between atmospheric water and both dynamical and mass fields, and the inherent non-Gaussian nature of atmospheric water variables[12–14]. Furthermore, GNSS observations over the ocean have gained significant attention in recent years, as reported by Ji et al.[15] and He et al.[16]. Nevertheless, establishing real ground-based GNSS stations on the sea remains challenging, leading to insufficient GNSS data availability. Consequently, the employment of idealized cases serves as an alternative approach to assess the influence of this observation type.

The objective of this study is to assess the impact of assimilating ground-based GNSS data within an idealized tropical cyclone scenario using an ensemble-based Kalman filter. This is achieved by integrating the Weather Research and Forecasting Model (WRF)[17] with the Parallel Data Assimilation Framework (PDAF; http://pdaf.awi.de, last access: August 1, 2024)[18] and performing and analyzing idealized assimilation experiments. Our goal is to gain insights into the potential benefits of utilizing ground-based GNSS data over the sea to enhance the accuracy and reliability of tropical cyclone predictions. For GNSS observations we focus on the online assimilation of PW and ZTD data, which are critical for improving tropical cyclone predictions. Compared to previous studies, we assimilate multiple observations, including temperature (T), horizontal wind components (U and V), and the additional variables PW and ZTD, using the localized error subspace transform ensemble Kalman filter (LESTKF)[19]. To implement this approach, we develop ground-based GNSS observation operators within the WRF-PDAF framework. These operators enable the seamless integration of GNSS data into the assimilation process. We then conduct a twin experiment over the ocean, based on a mesoscale idealized case of a tropical cyclone. By analyzing the assimilation results, we expect to gain valuable insights into the impact of ground-based GNSS data on tropical cyclone predictions. This study contributes to the ongoing efforts to improve the accuracy and reliability of tropical cyclone forecasting systems, ultimately leading to better decision-making and mitigation strategies for communities at risk.

The remainder of the study is structured as follows. "Methodology" introduces ensemble filters and the observation operators for the DA. The setup and configuration of the DA system are outlined in "Setup of data assimilation program". "Experimental design" discusses details of the experimental design for the idealized case studies. "Results and analysis" examines the parallel performance of the DA system build by coupling WRF and PDAF, the assimilation behavior of an example application with WRF. Finally, conclusions are drawn in "Discussion and conclusions".

Methodology

Ensemble Kalman filters (EnKFs, see e.g., Vetra-Carvalho et al.[20]) are data assimilation methods that combine the information from an ensemble of model states with observations to update the model state variables. In EnKFs, ensemble members are generated by perturbing the model initial conditions, and the assimilation is performed by computing analysis increments based on the ensemble spread and the observation-model misfit. Here, ensemble spread is the ensemble standard deviation (STD), which provides a measure of the distribution of the ensemble members around the ensemble mean. The analysis increments are subsequently added to the ensemble members to obtain the updated state variables. EnKF variants are particularly suitable for assimilating GNSS data due to their ability to handle non-linear dynamics of atmospheric models, like the LETKF[21] and the LESTKF [28, and also non-Gaussian distributions, like the NETF[22] and the LKNETF[23].

In this section, we introduce the WRF-PDAF model, the GNSS data for DA, the LESTKF assimilation scheme, and the observation operators.

WRF-PDAF

The WRF model is a widely used numerical weather prediction system, providing a versatile platform for simulating a broad spectrum of atmospheric processes suitable for both regional and global weather simulations. Developed by Shao and Nerger[24], WRF-PDAF integrates WRF-ARW version 4.4.1 with PDAF version 2.0 to facilitate robust data assimilation. This integration enables the incorporation of profile data into WRF, enhancing its initial conditions and contributing to improved forecast accuracy. The online coupling strategy of WRF-PDAF, in which PDAF is directly coupled to WRF, utilizes a fully parallel structure for data assimilation. Here, the data assimilation program integrates all model states concurrently utilizing a sufficient number of processes and the data assimilation is performed without the need of restarting the model. This approach guarantees the model's consistent temporal advancement, resulting in highly efficient data assimilation.

In this study, the model setup is the three-dimensional equivalent of case considered by Rotunno and Emanuel[25]. The domain size is 3000 km × 3000 km × 25 km, containing 200 × 200 × 20 grid points with a horizontal grid spacing of 15 km and a vertical grid spacing of 1.25 km. The Kessler microphysics scheme and the YSU boundary-layer physics are employed, while radiation schemes are not utilized. A capped Newtonian

relaxation scheme is used on potential temperature[25] which is a crude approximation for longwave radiation. This scheme is useful for idealized studies of maximum tropical cyclone intensity. The simulation spans a period of six days, starting from September 1, at 00:00 UTC (010000) and ending on September 7, at 00:00 UTC (070,000). The model time step is set to 60 s.

To initialize the simulation, both initial and boundary conditions are required. For our idealized tropical cyclone case the initial horizontally homogeneous environment is specified via a sounding data. The initial state is motionless ($u = v = 0$) and horizontally homogeneous, except an analytic axisymmetric vortex in hydrostatic and gradient-wind balance is added. The lateral boundary conditions are periodic to facilitate the simulation process. The default setup may not be optimal for complicated diagnosis of precipitation. These parameters of the default setup are adjustable to accommodate various requirements and preferences. Shao and Nerger[26] applied WRF-PDAF to conduct assimilation experiments of temperature profiles at different densities. The main difference in this study is the additional assimilation of PW and ZTD data. Additionally, we have supplemented the experiments with single-point experiments.

GNSS Data for DA

GNSS data, such as PW and ZTD observations, provide information about atmospheric moisture and can be assimilated using EnKFs to improve the representation of moisture fields in the model[27,28]. GNSS signals are bent, attenuated and delayed both by the ionosphere and troposphere. The ionospheric delay can be mostly reduced by linear combination of double-frequency observations. The water vapor content is responsible for the "wet" delay in the troposphere. A prevalent approach involves mapping the GNSS signal in the zenith direction and integrating it over a specified time period to derive a vertical column of tropospheric delay above each station, commonly referred to as the ZTD[29]. GNSS signals transmit through the troposphere and the signal delays are caused. The observed ZTD can be split into two parts: Zenith Hydrostatic Delay (ZHD) and Zenith Wet Delay (ZWD). The ZHD is estimated with the Saastamoinen[30] formula. In real cases, Precipitable Water vapor (PW) is retrieved from the ZWD as follows:

$$PW = Q \cdot ZWD \tag{1}$$

$$ZWD = ZTD - ZHD \tag{2}$$

$$Q = \frac{10^6}{461.525\left(\frac{375463}{T_m}+22.9726\right)} \tag{3}$$

$$T_m = \frac{\sum_{l=lts}^{l=lte} \frac{e_l}{T_l} \Delta h_l}{\sum_{l=lts}^{l=lte} \frac{e_l}{T_l^2} \Delta h_l} \tag{4}$$

$$ZHD = \frac{0.0022767P}{1-0.00266\cos(2\varphi)-0.00000029h} \tag{5}$$

Here Q is the proportionality factor. T_m denotes the vertical weighted mean temperature (in K) of the atmosphere. e_l, T_l and Δh_l denote the average vapor pressure (in hPa), average temperature (in K) and the thickness of the atmosphere at the $l-$ th layer, respectively. l is the layer index, ranging from the bottom layer lts to the top layer lte, specifically depending on the datasets used, such as ERA5[29,31]. P, φ and h are pressure, latitude and height of the station, respectively.

The GNSS data originates from ground-based stations, with real-time ZTD and PW data available on an hourly basis from these stations. Both the station's geographical location and the temporal resolution of the ZTD and PW data have to taking into account. For the data assimilation, the synthetic PW and ZTD observations are calculated hourly by the observation operators for PW and ZTD, respectively, acting on different model fields, as described in "PW and ZTD Observation operators". The two-dimensional ZTD/PW observations are positioned on all of the horizontal grid points. Synthetic U, V, and T observations represent sounding profile observations. In terms of profile data, the operators for T, U, and V directly operator on the model grid locations without any interpolations. Each profile consists of a vertical column of observations of T, U, and V located on grid points. It is usually impossible in the real scenario, even on land. However, this is precisely the purpose of our implementation of OSSE. We want to understand how data assimilation performs under this assumption.

LESTKF

The LESTKF has been applied in different studies to assimilate satellite data into atmosphere models[32], ocean models[33], atmosphere–ocean coupled models[34,35] and hydrological models[36]. The LESTKF is an efficient formulation of the EnKF, reviewed here to be able to discuss the particularities of the DA with respect to the ensemble filter. The analysis Eqs. (6)–(13) transform the forecast ensemble X^f of N_e model states into the analysis ensemble X^a:

$$X^a = X^f \left(w 1_{N_e}^T + \widetilde{W}\right) + \widetilde{x}^f 1_{N_e}^T \tag{6}$$

$$w = TA\left(HX^fT\right)^T R^{-1}(y - H\widetilde{x}^f) \tag{7}$$

$$\widetilde{W} = \sqrt{N_e - 1} T A^{1/2} T^T \tag{8}$$

$$A^{-1} = \alpha\,(N_e - 1)\,I + \left(HX^fT\right)^T R^{-1} HX^fT \tag{9}$$

$$T_{j,i} = \begin{cases} 1 - \frac{1}{N_e}\frac{1}{\frac{1}{\sqrt{N_e}}+1} for i = j, j < N_e \\ -\frac{1}{N_e}\frac{1}{\frac{1}{\sqrt{N_e}}+1} for i \neq j, j < N_e \\ -\frac{1}{\sqrt{N_e}} for i = N_e \end{cases} \tag{10}$$

$$USU^T = A^{-1} \tag{11}$$

$$A = US^{-1}U^T \tag{12}$$

$$A^{1/2} = US^{-1/2}U^T \tag{13}$$

Here, $\widetilde{x}^f$ is the forecast ensemble mean state and $1_{N_e}^T$ is the transpose of a vector of size N_e holding the value one in all elements. w is a vector of size N_e, which transforms the ensemble mean and $\widetilde{W}$ is a matrix of size $N_e \times N_e$, which transforms the ensemble perturbations. T is a projection matrix into the error subspace with $j = N_e$ rows and $i = N_e - 1$ columns. H is the observation operator. R is the observation error covariance matrix. A is a transform matrix in the error subspace. α with $0 < \alpha \leq 1$ is the forgetting factor[37]. U and S are the matrices of eigenvectors and eigenvalues, computed from the eigenvalue decomposition of A^{-1}.

A local analysis is performed by updating the model fields at each grid point of the model independently. Only observations within horizontal and vertical localization radii are considered when updating a grid point. Consequently, the observation operator is local and computes an observation vector within the influence radius based on the global model state. Additionally, each observation is weighted according to its distance from the grid point[21]. The localization weight for the observations is computed using a fifth-order polynomial with a shape resembling a Gaussian function[38]. The weighting is applied to the matrix R^{-1} in Eqs. (7) and (9). So, the localization process results in individual transformation weights w and $\widetilde{W}$ for each local analysis domain.

PW and ZTD Observation operators

Observation operators are used to transform model variables into observation space, thus computing the model equivalent to the actual observation. Specifically related to GNSS data the operator for PW is

$$PW\,(i,j) = \sum_{k=kts}^{k=kte} (\rho\,(i,j,k)\,q\,(i,j,k)\,\Delta h_k) \tag{14}$$

where i, j are horizontal model node indices. k is the index of the model layer, and kts=1 and kte=20 are the bottom layer and the top layer index as defined in "WRF-PDAF", respectively. ρ is air density (kg/m3), q is specific humidity (1), and Δh is the height difference between two consecutive model layers (m).

The observation operator for ZTD is

$$\begin{aligned} ZTD\,(i,j) = \frac{0.0022767 p_{sfc}(i,j)}{1-0.00266 cos(2\varphi)-0.00000029 h_{sfc}(i,j)} \\ +\sum_{k=kts}^{k=kte} \left(\frac{2.2110^{-7} p(i,j,k) q(i,j,k)}{t(i,j,k)} + \frac{3.7310^{-3} p(i,j,k) q(i,j,k)}{t^2(i,j,k)} \right) \frac{\Delta h_k}{0.622}. \end{aligned} \tag{15}$$

Here, h_{sfc} is the height of the model surface (m), p is pressure (Pa) and t is temperature (K). q, t and Δh are calculated from the model fields of WRF as follows:

$$q = \frac{qv}{1+qv} \tag{16}$$

$$t = (th + 300) \times \left(\frac{p}{100000}\right)^{0.286} \tag{17}$$

$$\Delta h = \Delta \left(\frac{ph+phb}{9.81}\right) \tag{18}$$

The model fields used in these functions are the perturbation geopotential (ph, m2/s), perturbation potential temperature (th, K), water vapor mixing ratio (qv, kg/kg), perturbation pressure (p, Pa), and base-state geopotential (phb, m2/s).

The observation operators of PW and ZTD are constructed based on the traditional approach. Our contribution lies in more explicitly formulating the equations within the WRF-PDAF system, thereby enabling accurate implementation. In one hand, since the operators of ZTD and PW are different, the results of the data assimilation should not be the same. Comparing the two different results is meaningful. In another hand, assimilating PW and ZTD should yield similar performance, which can be used to demonstrate the correctness of the construction of the observation operator and the assimilation process, making the results convincing.

Setup of data assimilation program

To enable the data assimilation, PDAF is coupled into the existing WRF framework. This coupling allows for the assimilation of GNSS data into WRF to improve its initial conditions and subsequently enhance its forecasts.

PDAF is a freely available open-source software developed to facilitate the implementation and application of ensemble and variational DA methods. It offers a generic framework that includes fully implemented and parallelized ensemble filter algorithms such as the LETKF, LESTKF, NETF, and LKNETF, along with related smoothers. PDAF provides functionality for adapting the model parallelization for parallel ensemble forecasts and includes routines for parallel communication between the model and filters. Like many large-scale

geoscientific simulation models, PDAF is implemented in Fortran and parallelized using the Message Passing Interface (MPI) standard[39] and OpenMP[40], ensuring optimal compatibility with such models. It can also be used with models implemented in other programming languages such as C and Python.

The online coupling strategy for DA is selected here utilizing the fully parallel structure. For this implementation, the time stepping for all ensemble states are is computed concurrently utilizing a sufficient number of processes on a compute cluster. With this, each model task integrates only one model state and the model is always going forward in time.

In this study, all of the variables needed by PDAF are inserted from WRF into the state vector. There are the x-wind component (u, m/s), y-wind component (v, m/s), z-wind component (w, m/s), perturbation geopotential (ph, m2/s), perturbation potential temperature (th, K), Water vapor mixing ratio (qv, kg/kg), Cloud water mixing ratio (qc, kg/kg), Rain water mixing ratio (qr, kg/kg), Ice mixing ratio (qi, kg/kg), Snow mixing ratio (qs, kg/kg), Graupel mixing ratio (qg, kg/kg), perturbation pressure (p, Pa), density (ρ, kg/m3) and base-state geopotential (phb, m2/s). Note that the variables p, ρ and phb are only used by the observation operators, but will not be updated by PDAF. So, only the rest of the variables will be updated and returned to WRF.

Experimental design

In the ideal cases, synthetic observations are used and generated from the model variables via observation operators. In this study, all synthetic observations were generated by adding Gaussian errors directly at the grid points without any interpolations. The standard deviations of the Gaussian errors were set to 1.2 K, 1.4 m/s, 1.4 m/s, 1 cm, and 4 cm for T, U, V, PW, and ZTD, respectively, following Bao and Zhang[41], 32.Pawel et al.[42] and Li et al.[43]. Therefore, the conventional observation operators, including U, V, and T, are just acting on the location of the model grid. For PW and ZTD, the model state variables are transformed into the observation space using the appropriate GNSS observation operators introduced in "PW and ZTD Observation operators". The PW and ZTD observations are then assimilated into the WRF model using the LESTKF.

In this section, the details of the ideal tropical cyclone case, the design of a single point experiment, and the experimental design of GNSS DA are described.

The tropical cyclone case

Tropical cyclones, also known as hurricanes or typhoons depending on the region, are powerful and destructive weather phenomena that form over warm ocean waters near the equator. These intense storms derive their energy from the latent heat released when moist air rises and condenses into clouds and precipitation. The Coriolis effect causes the storm to spin, with the direction of rotation determined by the hemisphere in which the cyclone forms. Tropical cyclones can have devastating impacts on coastal communities and infrastructure. Forecasting and monitoring tropical cyclones are essential for mitigating their impacts.

The test case used here is the idealized tropical cyclone case provided by WRF, which serves as a simplified representation of real-world atmospheric conditions. It provides a controlled environment for evaluating the performance of data assimilation methods utilizing identical twin experiments. This test case here we use is the same with Shao and Nerger[24], where one can find more details about the idealized tropical cyclone.

The atmospheric state variables, such as temperature, humidity, and wind fields from a forward run of the model create the known true state for comparison with assimilation results. This truth is used to generate synthetic observations. A control state is generated separately for the period September 3, at 12:00 UTC (031,200) to September 7, at 00:00 UTC (070,000) using the same initial fields as the truth. Therefore, the control state and true state are identical in all aspects except for their respective start times. The control simulation provides initial state estimate for the data assimilation. The flowchart of the cases is shown as Fig. 1.

Synthetic observations were generated hourly from the true state starting from 040,800 and ending at 051,400. The observations were generated for both single-point experiments and cycled DA experiments assimilating multiple variables. For the single-point experiments, only one set of observations of U, V, and T was generated at 040,800. These observations were located at the horizontal center of the model domain and vertical level 5,

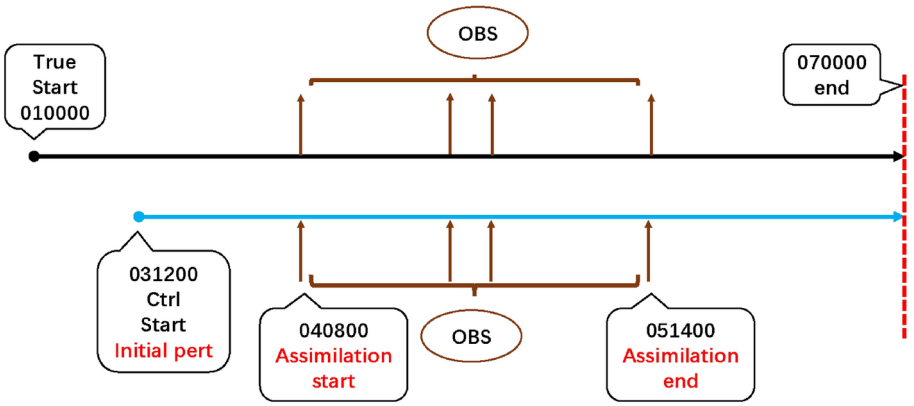

Fig. 1. The flowchart of the twin experiments. The black line represents the true state, and the blue line represents the control state.

Exp	Assimilated obs	offset	Updated vars	Localization(km)
1	T	1 K	U, V, T	800, 150, 50
2	U	1 m/s	U, V, T	800, 150, 50
3	V	1 m/s	U, V, T	800, 150, 50

Table 1. The design for single-point experiments.

Exp	Name	Member(s)	Assimilated obs	DA-Cycles
4	True	1	–	–
5	CTRL	1	–	–
6	ENS	40	–	–
7	daUVT	40	U, V, T	30
8	daPW	40	PW	30
9	daZTD	40	ZTD	30
10	daUVTPW	40	U, V, T, PW	30
11	daUVTZTD	40	U, V, T, ZTD	30

Table 2. The design for DA experiments.

corresponding to a height of 10 km. In the cycled DA experiments assimilating multiple variables, a total of 30 hourly observations of U, V, T, PW, and ZTD were generated. The U, V, and T observations were generated at all grid points in the model domain. On the other hand, PW and ZTD observations were generated from each vertical column of the model using the observation operators. In addition, Gaussian noise with standard deviation as described on "Experimental design". The generated observations are free of bias.

For the twin experiments, an initial perturbation is added to the control state at 031,200 to generate 40 ensemble members. The ensemble is spun up for 20 h. Subsequently, in the cycled DA, the observations are assimilated hourly into the ensemble during the analysis period from 040,800 to 051,400. Finally, an ensemble forecast is run without further assimilation from 051,400 until 070,000.

Single-point experiments

The single point experiments focus on assimilating observations at a specific location within the model domain. The design involves selecting a grid point of interest and assimilating observations at that point. These experiments allow for a detailed assessment of the assimilation impact on the model state variables at a specific location. Here, T is used to denote potential temperature th. As depicted in Table 1, the assimilation of a single observation of T, U and V located at the specific location was carried out in three separate experiments, namely exp.1, exp.2 and exp.3. The observations T, U, and V have offsets of 1 k, 1 m/s, and 1 m/s, respectively, relative to the control fields. These observations were assimilated to compute a multivariate update of U, V and T. To determine the optimal localization distance, the different horizontal distances 800 km, 150 km, and 50 km are tested in each experiment. Localization distances of 200 km and 100 km were also tested for tuning, but not shown here. The vertical localization radius is identical in all cases matching the height of the model top. The purpose of these tests was to select the localization distance that yielded the best results. To facilitate analysis and verification, there is no radius for PW or ZTD.

In this study, a forgetting factor of $\alpha = 0.97$ was used in Eq. (9). The forgetting factor is a scaling parameter applied to the ensemble spread in order to avoid underestimation of the forecast uncertainty. The ensemble variance is inflated by $1/\alpha$. The forgetting factor was determined based on the ensemble spread, which reflects the variability or uncertainty within the ensemble members. By appropriately adjusting the forgetting factor and setting observation errors, the assimilation process can effectively incorporate the available information from observations and ensemble members, resulting in improved forecast accuracy and reliability.

Experimental design for cycled DA

The experimental design for DA with multiple observations involves assimilating synthetic conventional and GNSS observations into the WRF model. The GNSS DA experiment aims to enhance the representation of moisture fields through the integration of GNSS observations. This assimilation process aims to utilize the precision PW and ZTD data to refine and correct the model predictions of humidity and other related atmospheric variables. The ultimate objective is to achieve a more accurate representation of moisture fields, thereby enhancing the overall accuracy and reliability of weather predictions. The impact of assimilating these observations on the model representation of atmospheric moisture is evaluated through a comparison between the assimilated and true states. By conducting these experiments on an idealized case, the performance and effectiveness of WRF-PDAF in assimilating observations and improving the model representation of atmospheric variables can be evaluated.

Table 2 provides an overview of the experiments performed here. Two single runs were used to generate the true state (Exp. 4, 'True') and control state (Exp. 5, 'CTRL'), as described in "The tropical cyclone case". These distinct states served as the basis for further analysis and experimentation in the study. To generate the initial

ensemble, perturbations were generated using second-order exact sampling[37] from the model variability of hourly snapshots from 010000 to 031,200. These perturbations were added to the control state at 031,200 to generate an ensemble of 40 states. Subsequently, a free ensemble run of the 40 members (Exp. 6, 'ENS') was conducted. The purpose of this ensemble run was to generate a collection of model states that encompassed a range of possible variations and uncertainties. The same initial ensemble members were utilized in the assimilation experiments. Starting from the initial ensemble, assimilation experiments were conducted over 30 analysis cycles. Different experiments assimilating the conventional observations U, V, T, or separately the GNSS observations PW or ZTD, as listed in Table 2, were performed. In addition, the experiments 10 and 11 assimilated a combination of direct observations alongside with GNSS observations. PW and ZTD are assimilated separately to assess how far these observations have different effects. This integration leverages the complementary nature of the two datasets. These different assimilation experiments were carried out to evaluate the impact of assimilating specific types of observations on the model state.

Results and analysis

Single-point experiments

Figure 2 shows the increments resulting from the single-point assimilation experiments detailed in Table 1. Note that each assimilation of T, U, and V observations can affect all of the U, V, and T model fields through the multivariate DA update. In contrast to the isotropic increments of 3DVAR and 4DVAR, the increments used in LESTKF are anisotropic due to the flow-dependent features of the background error covariance.

If there is no localization, the increments will be distributed throughout the entire simulation region. However, increments far from the observation are generally unreliable, and the correlations between the observation point

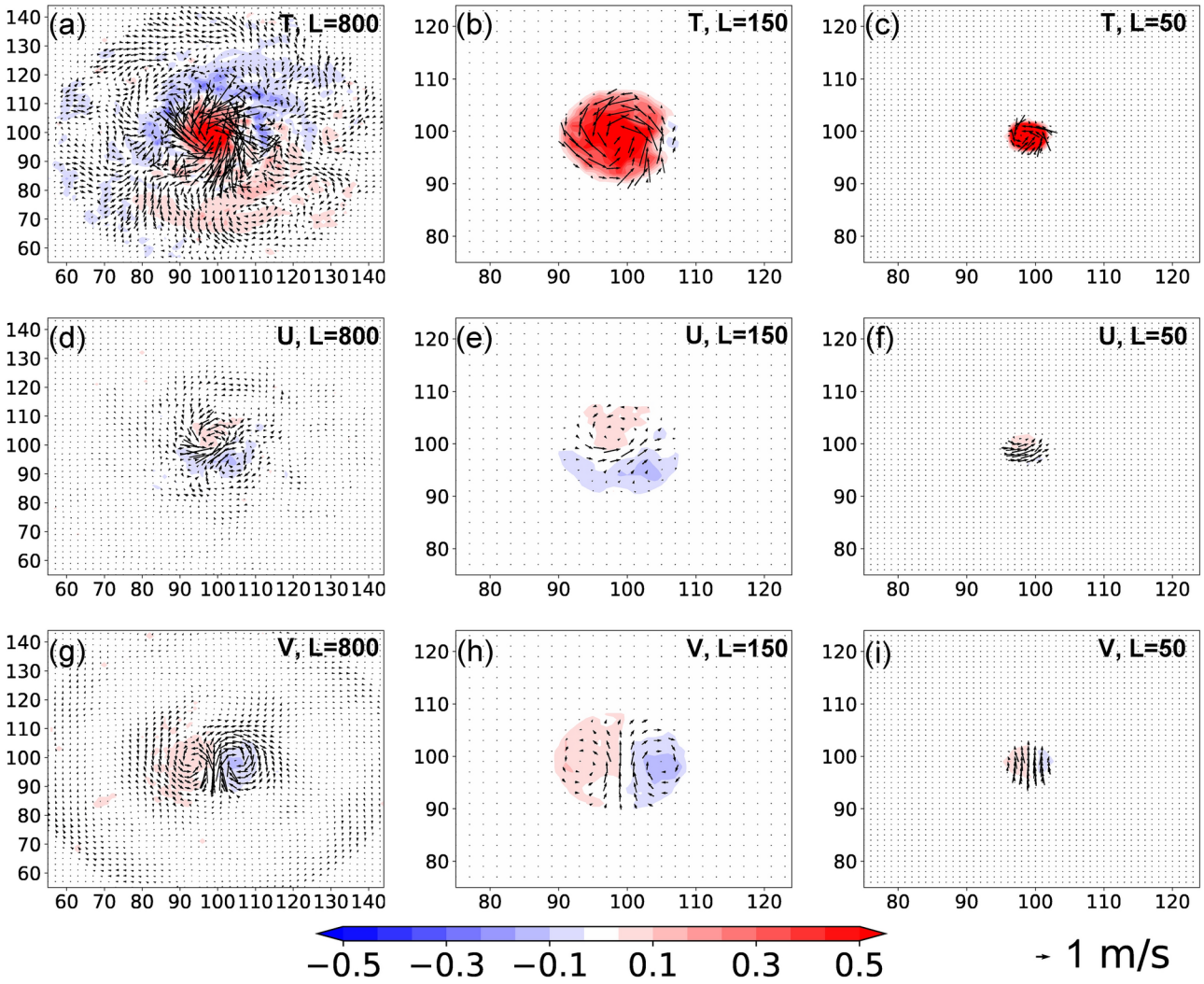

Fig. 2. The spatial distribution of the T, U and V increments of the single-point experiments at 031,200 with different localization distances 800 km, 150 km, and 50 km ((**a**–**c**: results of exp. 1 assimilating T; (**d**–**f**): results of exp. 2 assimilating U; (**g**–**i**)): results of exp. 3 assimilating V). The shade represents the T increments (K), while the arrows represent the wind velocity (combined U and V) increments (m/s).

and distant grid points were considered spurious. To address this concern, selecting an appropriate localization distance becomes crucial. Past research often made such choices or even developed adaptive schemes based on the root mean square error (RMSE). However, in our study, since the dynamics are known in the ideal case, we aim to determine the localization distance from the dynamic perspective. A well-suited localization distance should accurately reflect the relationships between temperature and wind while also avoiding spurious correlations. When the localization distance was set to 800 km, the region with spurious correlations reduced compared to using no localization, but some areas with spurious increments remained (Fig. 2, column (1). When the localization distance was further reduced to 150 km, the increments were only distributed closely around the single observation point (Fig. 2, column (2). The fifth-order polynomial mentioned in "LESTKF" resulted in decreasing increments as the distance from the observation point increases. Moreover, positive T increments caused cyclonic-type wind increments, while negative T increments caused anticyclonic-type wind increments, consistent with the gradient-wind balance. The localization distance of 50 km led to even smaller areas of increments around the single observation point (Fig. 2, column (3). However, the area of increments was too limited to clearly observe the relationship of the gradient-wind balance, especially in Fig. 2f,i. Despite the reduced spurious correlations, the extremely localized increments hindered the ability to capture the meso-scale flow patterns and relationships. Based on the results and observations provided, a localization distance of 150 km was chosen as the most suitable for the assimilation experiments in this study.

Cycled GNSS DA

In Fig. 3a, the RMSE of specific humidity (Qv) from the different experiments listed in Table 2 is displayed. The RMSE of the ensemble forecast (ENS) is lower than that of the control run from the true state (True). This means that the ensemble members generated using second-order exact sampling represent the range of possible atmospheric states and the ensemble mean properly represents the most likely forecast. The RMSE when assimilating U, V, T data (UVT) is lower than that of ENS, indicating that the assimilation process improves the accuracy of the model prediction. The RMSEs from the experiments daPW and daZTD appear to be similar, with the RMSE of daZTD is slightly lower than daPW. The RMSEs from the experiments daUVT, daUVTPW, and daUVTZTD are similar. However, the RMSEs from daUVTPW and daUVTZTD are lower than that of daUVT. Among all the experiments, the lowest RMSE is observed in daUVTZTD.

In Fig. 3b, the STD of the ensemble of Qv is shown for the different experiments. The STD provides an estimate of the uncertainty in the state estimate. The STD of the experiment daPW is slightly lower than that of ENS. The STD of experiment daZTD is lower than that of daPW. This suggests that the assimilation of either PW or ZTD data has helped to reduce the uncertainty among the ensemble members, leading to a more consistent forecast. The experiments daUVT, daUVTPW and daUVTZTD have almost the same STD, which is lower than the others. This suggests that the assimilation of conventional data and of multiple observations (U, V, T, PW/ZTD) in these experiments have led to a similar reduction in the spread of specific humidity among the

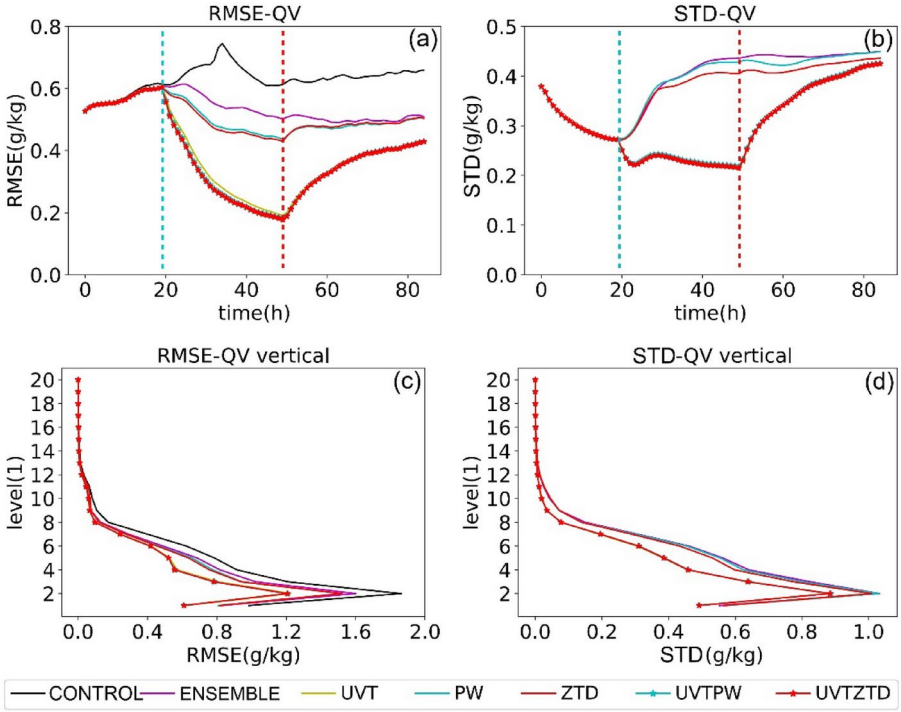

Fig. 3. Upper row: Time series of RMSE (**a**) and STD (**b**) of Qv from 031,200 to 070,000. Lower row: Vertical profile of time-average of Qv RMSE (**c**) and STD (**d**). The blue dotted lines in (**a**) and (**b**) show the start time of the DA process, while the red dotted lines represent the its endpoint).

ensemble members, contributing to a more constrained forecast. The pattern of the STDs is similar to the RMSEs during the analysis period, indicating, as expected, that the ensemble STD is influenced by the assimilation process. Lower RMSEs correspond to lower STDs. An approach to evaluate the capability of an ensemble system in quantifying prediction uncertainty is by examining the relationship between the spread among the forecasts of individual ensemble members and the skill of their mean forecast, known as the spread-skill relationship[44]. Several methods exist to quantify this relationship. Talagrand[45] argued that a statistically consistent ensemble should have an average STD matching the RMSE of its mean forecast. We observe that, indeed, the STD and RMSE generally correspond quite well. However, all of the STDs become closer during forecast period, especially at the later time of the experiment. In Fig. 3c,d, it can be observed that the decreases in RMSEs and STDs of Qv are primarily seen at the middle and low levels (below level 12). This suggests that the assimilation process has a more significant impact on improving the accuracy and reducing the variability of Qv at these levels. However, the results of T, U, and V show the decreases in RMSEs and STDs of these variables at all levels (figures omitted). Figure 3 provides insights into the performance of different data assimilation experiments in improving the accuracy and reducing the STD of Qv, T, U, and V variables during the specified time period (from 031,200 to 070,000).

With the aid of flow-dependent cross-variable background error covariances, the assimilation of U, V, and T yields Qv corrections, resulting in an improved state compared to assimilating PW or ZTD alone. This could be attributed to the nature of the observations themselves. The U, V, and T observations are direct measurements and represent three-dimensional variables, providing a comprehensive and detailed information about the atmospheric conditions. On the other hand, the PW and ZTD observations are indirect two-dimensional data, which may have some limitations in capturing the complete atmospheric state. The direct and three-dimensional nature of U, V, and T observations likely contributes to their larger impact on the assimilation process and the resulting improvements in the state. Furthermore, the assimilation of GNSS data generates slight wind and temperature corrections through the same flow-dependent mechanism (figures omitted). Additionally, the assimilation of multiple data types (U, V, T, PW/ZTD) contributes to an enhanced initial cyclone circulation. This improvement can be credited to the assimilation of diverse data types, which effectively corrects the temperature, wind, and Qv fields. This indicates that despite the significant improvements achieved through conventional observations, the inclusion of GNSS observations can offer additional valuable information, leading to further enhancements in cyclone simulation.

In the idealized case, the primary distinction between PW and ZTD stems from the different observation operators outlined in Eqs. (14–15). Simultaneously, due to varying observational errors, the RMSE exhibits different performances. This discrepancy is also reflected in the lower RMSEs and STDs of U, V, and T from daUVTZTD compared to daUVTPW. In previous real case studies[14,46], opinions vary on whether assimilating PW or ZTD yields better results. From our perspective, the superiority of either depends on the quality of the data itself. In real cases, ZTD is derived first, followed by the derivation of PW from ZTD. It is crucial to note that the value of PW does not solely depend on the ZTD but is also influenced by the additional variables (p, t) in Eqs. (1–5). If the quality of ZTD surpasses that of p and t, the quality of PW may be inferior to that of ZTD, and vice versa.

To assess the estimate model fields, we show the ensemble means for the ensemble experiments as it is common practice in ensemble DA. The spatial distribution of T, U, and V at the 850mb level at the initial time (031,200) are shown in Fig. 4. Figure 4a represents the true state, showing the actual distribution of T, U, and V. Figure 4b represents the control run, which is similar to the ensemble run (Fig. 4c), but both differ significantly from the true state. Figure 4d displays the difference obtained by subtracting the true state from the ensemble mean. The differences of T are positive in the outer areas but negative in the central region, while most wind velocities exhibit an anticyclonic pattern. As a result of the distinct start times, the cyclone in the true state has progressed for 60 h, whereas the control state's cyclone remains at its earlier stage. During the development of the cyclone, the temperature in the central region increases, while it decreases in the outer areas. Concurrently, the wind field intensifies over time. This phenomenon can be attributed to the interplay between thermodynamics and dynamics within the cyclonic system.

Given its lowest RMSE, the daUVTZTD experiment is selected for further comparative analysis in this study. At this first analysis step of the DA process the analysis state gets closer to the true state compared to CTRL and ENS. The misfit between daUVTZTD and True is smaller than that between ENS and True at the initial time (figures omitted). However, the difference is larger compared to the final assimilation time. The larger error after the first analysis is mainly due to the substantial magnitude of the prescribed observation errors. Thus, the impact of the observations may not be immediately evident or prominent. However, by incorporating model observational information over time, the state estimate is gradually improved.

Figure 5 represents the 30th DA cycle and final assimilation time, which is 50 h after the start time of control run. In the control run (Fig. 5b), T is lower, and the cyclone is weaker than in the true state (Fig. 5a). In ENS (Fig. 5c), T is higher than in the true state, whereas the cyclonic circulation remains weaker. An evident underestimation of temperature is observed at the cyclone center, whereas the temperature is overestimated in the areas outside the cyclone edge. The analysis state (Fig. 5d) is closest to the true state. The overall DA-induced change in the model state, depicted in Fig. 5e, demonstrate the impact of DA. The improvements of T are predominantly concentrated at the center of the cyclone and the surrounding area outside the edge of the cyclone. The differences between the analysis and the true state (Fig. 5f) are very small across the model domain.

Next to the effect on the temperature and velocity fields, we assess the effect of the DA on Qv in Fig. 6. At 051,400, Qv in the control run (Fig. 6b) appears to be higher than that in the true state (Fig. 6a) across the entire region. In contrast, the ensemble run (Fig. 6c) shows a lower Qv compared to the control run, yet it lacks accuracy in simulating the cyclone pattern around its center. Similar to the temperature, the Qv distribution of the analysis state (Fig. 6d) is the closest to the true state. The DA-induced change, depicted in Fig. 6e, illustrates

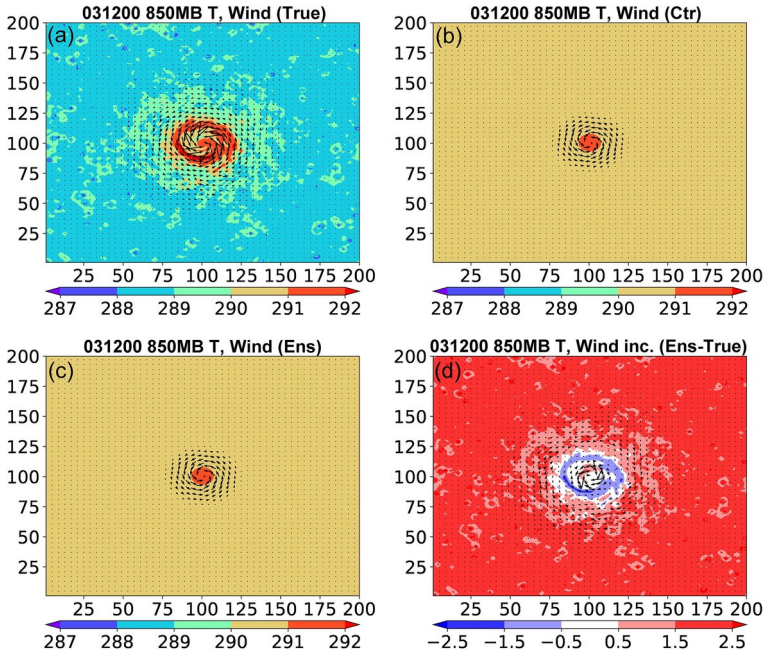

Fig. 4. Spatial distribution of T, U, and V at the 850 mb level at initial time 031,200 of the control run (single state for True and Ctr; ensemble mean for the ensemble experiment). The shade represents the temperature (K) distribution, while the arrows represent the wind velocity (m/s). 4 (**a**): True; 4 (**b**): CTRL; 4 (**c**): ENS; 4 (**d**): difference between ENS and True.

the impact of the DA on the Qv field, with adjustments evident throughout the domain but predominantly concentrated at the cyclone's center. The resulting misfits between the analysis state and the true state (Fig. 6f) are generally very small across the entire simulation region.

As the assimilation cycles progress, an increasing amount of information is assimilated into the background field. With more observations being incorporated, the analysis field progressively approaches the true field. At the 30th assimilation cycle, all available observations have been assimilated, resulting in the analysis field being the closest approximation to the true field (see Fig. 5). After a subsequent 20 h forecast without DA, the T and Qv patterns of the control run are significantly different from the true state, and the cyclone is still weaker than in the true state. T of the ensemble run near the center is lower than the true state, and the cyclone remains weaker than that in the true state. The analysis state is still closer to the true state than CTRL and ENS, but the misfits between daUVTZTD and the true state are larger than those at the time of final assimilation (figures omitted). In the absence of observation constraints, the simulated values in the assimilation experiments gradually deviate from the real values. However, despite this deviation, the assimilation experiments consistently outperformed the control state over time. For the limited spread of analysis ensemble, the 40 ensemble realizations for the daUVTZTD show similar behavior. It is worth noting that at 051,400, several isolated points emerge in both the T (Fig. 5f) and the Qv (Fig. 6f) fields, particularly in the surrounding area outside the cyclone edge.

In addition, we assess the impact of the DA on rainfall, focusing on the 24-h accumulated precipitation. At 051,400, the maximum precipitation in the control run (Fig. 7b) is less than that of the true state (Fig. 7a). The patterns of their distribution are notably distinct. The ensemble run (Fig. 7c) exhibits an even lower rainfall level than the control run, and it continues to miss the cyclonic pattern centered around its core. In line with the findings for other variables, the distribution of rainfall in the analysis state (Fig. 7d), aligns most closely with the true state.

Table 3 shows that the mean RMSEs of Qv, T, U, and V in all vertical levels, as well as 24 h rainfall for the ensemble forecast (ENS) are similar to those of the single control run (CTRL) at the initial time. At the 1st DA cycle, RMSEs for the ensemble forecast are smaller than those of the control run, and the RMSEs for the analysis are smaller than those for the ensemble run. At the 30st DA cycle—the final assimilation time—the RMSEs for daUVTZTD are the smallest among all experiments and assimilation times. After 20-h free forecast, the RMSEs for the ensemble forecast are still smaller than those for the control run. The RMSEs for assimilation run daUVTZTD are smaller than those for ENS, but larger than those for the analysis at the 30st DA cycle. In contrast to the results from previous studies[47,48], the RMSEs in our study show significant reductions by the DA, primarily attributed to the inclusion of additional conventional data and a higher assimilation rate of GNSS data. These enhancements have collectively contributed to reducing the forecast errors and increasing the accuracy of our simulations.

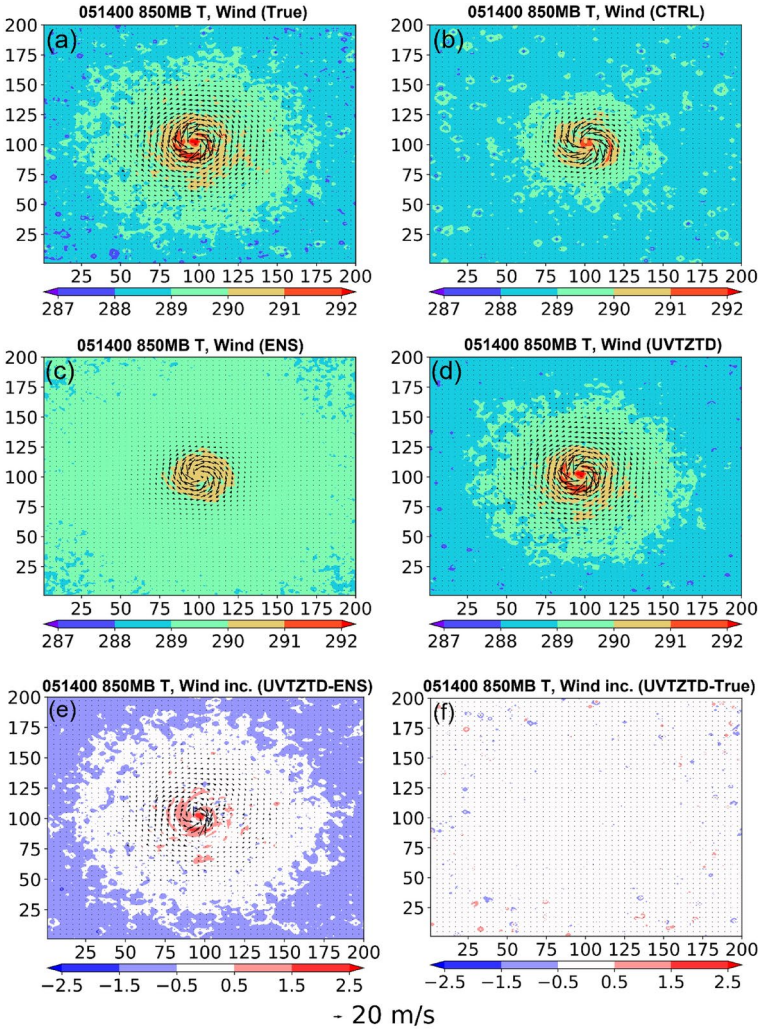

Fig. 5. T, U and V at level 850mb at 30th DA time 051,400. The shade represents the temperature (K) distribution, while the arrows represent the wind velocity (m/s). 5 (**a**): True state; 5 (**b**): CTRL; 5 (**c**): ENS; 5 (**d**): daUVTZTD; 5 (**e**): difference between daUVTZTD and ENS; 5 (**f**): difference between daUVTZTD and True.

Discussion and conclusions

In this study, a tropical cyclone twin experiment was conducted to evaluate the effect of assimilating conventional and GNSS data in different configurations. The assimilation results provide valuable insights into the performance of the ensemble Kalman filter LESTKF applied in WRF-PDAF, the developed GNSS operator, the impact of GNSS DA on the model forecast accuracy, and the behavior of the analyzed fields. A suitable localization distance needs to be selected to balance the assimilation impact with the preservation of meso-scale flow patterns. Specifically, the localization distance chosen for this study was determined based on the model dynamics, rather than solely relying on numerical values of the RMSE. This decision was made due to the evident correlation between temperature and wind in the idealized scenario, which provides a more physically meaningful basis for selecting the localization distance. However, the localization distance is case-dependent and not a general value. In practical applications, a typical localization radius of 1000 km is commonly used for global modeling and data assimilation systems[49]. However, for convective weather systems utilizing high-resolution models and observations, a much shorter radius of 10 km has been found to be more appropriate[50]. Nonetheless, experiments with real data conducted by Dong et al. (2011) suggest that a smaller localization radius is necessary to achieve better analysis accuracy with denser observing networks. Periáñez et al.[51] determined an optimal localization radius through heuristic arguments, assuming a uniform observing network, and also recommend using a smaller localization radius for denser observations. Kirchgessner et al.[52] proposed a scheme for adaptive localization without tuning. These studies indicate a potentially complex relationship between observing networks and localization radii. However, in real-world applications, the localization radius may be influenced by other factors. For instance, it is known that localization affects the balance in the model state, and a longer localization radius will have a smaller impact on the balance. Consequently, one might prefer a longer localization radius in multivariate assimilation applications. Additionally, when assimilating real observations, biases can

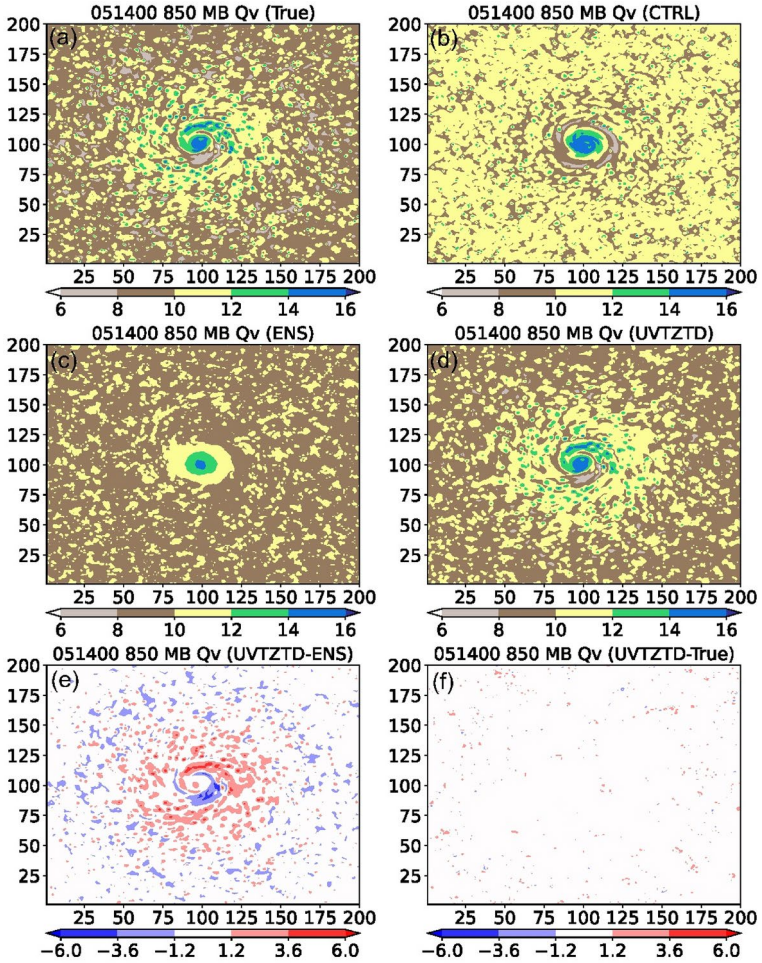

Fig. 6. The spatial distribution of Qv at level 850mb at 30th analysis time 051,400. The shade represents the distribution of Qv (g/kg), while the contours delineate the differences in Qv with an interval of 1 g/kg. Specifically, Fig. 6 (**a**) depicts the true state, 6 (**b**) shows CTRL, 6 (**c**) represents ENS, 6 (**d**) displays the results of daUVTZTD, 6 (**e**) illustrates the increments between the daUVTZTD and the ENS simulations, and 6 (f). depicts the misfits between the daUVTZTD and the true state.

occur, and the observation error covariance matrix might be inaccurately estimated. It remains unclear to what extent these factors necessitate adapting the localization radius to achieve overall optimal assimilation results. Therefore, tuning is still necessary. Perhaps, the effective spatial resolution[53] of the model[54] could be applied to determine the localization. Corresponding to "PW and ZTD Observation operators", assimilating PW and ZTD yields similar results. This proves that the construction of the assimilation operator and the implementation of the assimilation process are reliable. From another perspective, different operators of PW and ZTD caused differences in the DA performance. The DA results are influenced by the magnitude of the observation errors. In real cases, the superiority of either also depends on the data quality as described in "Cycled GNSS DA".

This study outperforms previous research[55–60] by achieving the most accurate assimilation results, evidenced by the lowest RMSEs and the most similar distributions with the true state. This superior performance can be attributed to the utilization of high-fidelity synthetic observations, which are not only precise but also have a high spatial resolution, characterized by a full 100% density on model grids. However, assimilating observations at lower density can still have a significant effect, at least for conventional observations as was shown by Shao and Nerger[26]. The analysis step significantly improves the accuracy of the model forecast compared to the control run or ensemble forecast. Assimilating the conventional observations U, V, and T, leads to increments that align with expected atmospheric features, such as cyclone patterns in this ideal case. Compared with previous studies, multiple observations, such as T, U, V, as well as PW and ZTD, which are derived from GNSS data, were assimilated using the LESTKF. This generally improved the forecast accuracy, compared to assimilating either conventional or GNSS data. The lower RMSEs compared to previous studies show the effectiveness of the applied assimilation method and the selected observed variables.

The key findings are significant as they contribute to the understanding of the impact of assimilating ground-based GNSS data on the forecast accuracy of tropical cyclone. They highlight the effectiveness of the assimilation process in improving the accuracy of the forecast and provide insights into the behavior of analyzed fields in a tropical cyclone. Additionally, the study identifies the benefits of assimilating multiple observation types.

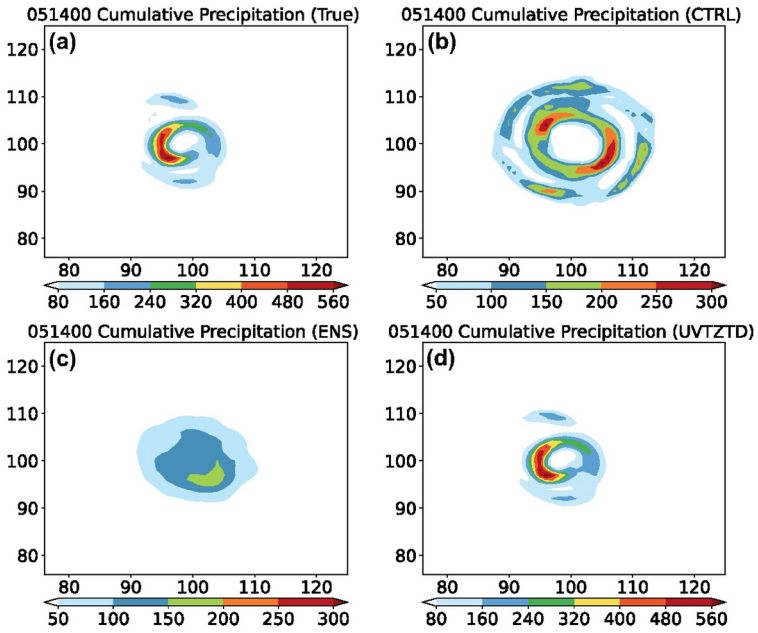

Fig. 7. The spatial distribution of 24h cumulative precipitation (in mm) at the 30th analysis time 051,400. Shown are the (**a**) true state, (**b**) CTRL, (c) ENS, (**d**) daUVTZTD.

Time	Experiment	Qv (g/kg)	U (m/s)	V (m/s)	T (K)	Rainfall (mm)
031,200	CTRL	0.534	1.922	1.859	1.195	0
	ENS	0.538	1.927	1.869	1.180	0
040,800 (Start of DA)	CTRL	0.612	2.155	2.125	0.951	16.708
	ENS	0.606	1.986	1.967	0.922	15.413
	daUVTZTD	**0.557**	**0.852**	**0.808**	**0.529**	**14.484**
051,400 (End of DA)	CTRL	0.622	1.909	2.075	1.096	17.165
	ENS	0.507	1.760	1.894	0.953	12.500
	daUVTZTD	**0.188**	**0.238**	**0.238**	**0.188**	**3.746**
061,000 (20 h forecast)	CTRL	0.639	1.912	2.245	1.162	22.648
	ENS	0.501	1.659	1.855	0.715	12.831
	daUVTZTD	**0.391**	**0.690**	**0.724**	**0.569**	**7.188**

Table 3. Mean RMSEs of Variables of experiments in all the vertical levels, as well as 24 h rainfall, at initial time 031,200, 1st DA cycle 040,800, 30st DA cycle 051,400 and 20 h forecast time 061,000. The minimum values for each element at each time point are in bold.

Assimilating ground-based GNSS data, such as PW and ZTD, offers several benefits in the tropical cyclone simulation. Ground-based GNSS data provide valuable information about atmospheric water vapor and can improve the representation of moisture fields in numerical weather prediction models. Assimilating ground-based GNSS data can hence improve the initialization of water vapor fields, and help capture mesoscale features related to atmospheric moisture. The findings of this study highlight the potential applications of assimilating ground-based GNSS data in improving weather forecasts in marine areas and demonstrate that it is essential to establish real ground-based GNSS observation stations over the ocean. By understanding the behavior of analyzed fields and the impact of assimilation, researchers and meteorologists can enhance forecast accuracy.

In conclusion, this research demonstrates the effectiveness of ground-based GNSS data assimilation using the ensemble Kalman filter LESTKF in improving tropical cyclone simulation accuracy. The findings emphasize the benefits of assimilating multiple observation types, and the potential applications of assimilating ground-based GNSS data. The construction of ground-based GNSS observation stations over the ocean is highly promising and essential. The utilization of the flow-dependent, cross-variable background error covariances in the LESTKF enables us to fully leverage the advantages of this data. By further advancing the LESTKF and incorporating GNSS operators in the data assimilation process, we can enhance simulation capabilities for tropical cyclones and have the opportunity to provide more accurate and reliable predictions for various applications, including network design, weather monitoring, disaster management, and climate studies. However, the study should acknowledge potential limitations, such as the use of an idealized twin experiment with synthetic observations.

Representation and model errors are not present here. The inherent non-Gaussian nature of atmospheric water variables are also not considered. Future research directions may involve investigating alternative data assimilation methods, in particular nonlinear methods, to address the limitations and challenges encountered. Investigating advanced techniques, such as adaptive localization or ensemble-based adaptive observation strategies, can potentially enhance the assimilation process.

Data availability

Dataset can be download at https://doi.org/https://doi.org/10.5281/zenodo.10335684.

Received: 28 May 2024; Accepted: 11 September 2024
Published online: 17 September 2024

References

1. Lorenc, A. C. Analysis methods for numerical weather prediction. *Q. J. R. Meteorol. Soc.* **112**(473), 1177–1194 (1986).
2. Song, L., Shen, F., Shao, C., Shu, A. & Zhu, L. Impacts of 3DEnVar-Based FY-3D MWHS-2 radiance assimilation on numerical simulations of landfalling typhoon ampil (2018). *Remote Sens.* **14**, 6037. https://doi.org/10.3390/rs14236037 (2022).
3. Risanto, C. B. *et al.* The impact of assimilating GPS precipitable water vapor in convective-permitting WRF-ARW on North American monsoon precipitation forecasts over Northwest Mexico. *Monthly Weather Rev.* **149**(9), 3013–3035 (2021).
4. Blewitt, G. V., Hammond, W. & Kreemer, C. Harnessing the GPS data explosion for interdisciplinary science. *Eos*https://doi.org/10.1029/2018EO104623 (2018).
5. Hdidou, F. Z. *et al.* Impact of the variational assimilation of ground-based GNSS zenith total delay into AROME-Morocco model. *Tellus A Dynamic Meteorol. Oceanogr.* **72**(1), 1–13. https://doi.org/10.1080/16000870.2019.1707854 (2020).
6. Torcasio, R. C. *et al.* The impact of GNSS Zenith Total Delay data assimilation on the short-term precipitable water vapor and precipitation forecast over Italy using the WRF model. Nat. Hazards Earth Syst. Sci. Dis. https://doi.org/10.5194/nhess-2023-18 (2023).
7. Singh, R., Ojha, S. P., Puviarasan, N. & Singh, V. Impact of GNSS signal delay assimilation on short range weather forecasts over the Indian region. *J. Geophys.Res.-Atmos.* **124**, 9855–9873. https://doi.org/10.1029/2019JD030866 (2019).
8. Giannaros, C. *et al.* Assessing the impact of GNSS ZTD data assimilation into the WRF modeling system during high-impact rainfall events over Greece. *Remote Sens.* **12**(3), 383. https://doi.org/10.3390/rs12030383 (2020).
9. Yang, S. C. *et al.* A case study on the impact of ensemble data assimilation with GNSS-Zenith total delay and radar data on heavy rainfall prediction. *Monthly Weather Rev.* **148**(3), 1075–1098 (2020).
10. Vaquero-Martínez, J. & Antón, M. Review on the role of GNSS meteorology in monitoring water vapor for atmospheric physics. *Remote Sens.* **13**(12), 2287 (2021).
11. Rohm, W., Yuan, Y., Biadeglgne, B., Zhang, K. & Marshall, J. L. Ground-based GNSS ZTD/IWV estimation system for numerical weather prediction in challenging weather conditions. *Atmospheric Res.* **138**, 414–426. https://doi.org/10.1016/j.atmosres.2013.11.026 (2014).
12. Bannister, R. N., Chipilski, H. G. & Martinez-Alvarado, O. Techniques and challenges in the assimilation of atmospheric water observations for numerical weather prediction towards convective scales. *Q.J. R. Meteorol. Soc.* **146**, 1–48. https://doi.org/10.1002/qj.3652 (2020).
13. Christophersen, H., Sippel, J., Aksoy, A. & Baker, N. L. Recent advancements for tropical cyclone data assimilation. *Ann. N. Y. Acad. Sci.* **1517**, 25–43. https://doi.org/10.1111/nyas.14873 (2022).
14. Christophersen, H., Ruston, B. & Baker, N. L. Assimilation of GNSS zenith total delay in NAVGEM. *J. Geophys. Res. Atmos.* https://doi.org/10.1029/2022JD037502 (2023).
15. Ji, S., Sun, Z., Weng, D., Chen, W. & He, K. High-precision ocean navigation with single set of beidou short-message device. *J. Geodesy* **93**(9), 1589–1602. https://doi.org/10.1007/s00190-019-01273-7 (2019).
16. He, Z., Chen, W., Yang, Y. & Shen, M. Sea target detection using the GNSS reflection signals. *GPS Solutions* https://doi.org/10.1007/s10291-023-01493-7 (2023).
17. Skamarock, W. C., Klemp, J. B., Dudhia, J., et al. A Description of the Advanced Research WRF Model Version 4.3 (No. NCAR/TN-556+STR). (2021).
18. Nerger, L. & Hiller, W. Software for Ensemble-based Data Assimilation Systems-Implementation Strategies and Scalability. *Comput. Geosci.* **55**, 110–118 (2013).
19. Nerger, L. *et al.* A unification of ensemble square root filters. *Monthly Weather Rev.* **140**, 2335–2345 (2012).
20. Vetra-Carvalho, S. et al. State-of-the-art stochastic data assimilation methods for high-dimensional non-Gaussian problems. *Tellus A* **70**(1), 1445364. https://doi.org/10.1080/16000870.2018.1445364 (2018).
21. Hunt, B. R., Kostelich, E. J. & Szunyogh, I. Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter. *Physica D Nonlinear Phenomena* **230**, 112–126 (2007).
22. Tödter, J. & Ahrens, B. A second-order exact ensemble square root filter for nonlinear data assimilation. *Monthly Weather Rev.* **143**, 1347–1467 (2015).
23. Nerger, L. Data assimilation for nonlinear systems with a hybrid nonlinear Kalman ensemble transform filter. *Q. J. R. Meteorol. Soc.* https://doi.org/10.1002/qj.4221 (2022).
24. Shao, C. & Nerger, L. WRF-PDAF v1.0: Implementation and application of an online localized ensemble data assimilation framework. *Geoscientific Model Dev.* **17**, 4433–4445. https://doi.org/10.5194/gmd-17-4433 (2024).
25. Rotunno, R. & Emanuel, K. A. An air-sea interaction theory for tropical cyclones. Part II. Evolutionary study using a nonhydrostatic axisymmetric numerical model. *J. Atmospheric Sci.* **44**, 542–561 (1987).
26. Shao, C. & Nerger, L. The impact of profiles data assimilation on an ideal tropical cyclone case. *Remote Sens.* **16**, 430. https://doi.org/10.3390/rs16020430 (2024).
27. Bennitt, G. V. & Jupp, A. Operational assimilation of GPS zenith total delay observations into the Met Office numerical weather prediction models. *Monthly Weather Rev.* **140**(8), 2706–2719 (2012).
28. Mascitelli, A. *et al.* Assimilation of GPS Zenith Total Delay estimates in RAMS NWP model: Impact studies over central Italy. *Adv. Space Res.* **68**(12), 4783–4793 (2021).
29. Wagner, A., Fersch, B., Yuan, P., Rummler, T. & Kunstmann, H. Assimilation of GNSS and Synoptic Data in a Convection Permitting Limited Area Model: Improvement of Simulated Tropospheric Water Vapor Content. *Front. Earth Sci.* https://doi.org/10.3389/feart.2022.869504 (2022).
30. Saastamoinen, J. Contributions to the theory of atmospheric refraction. *Bull. Geodesique* **105**, 279–298. https://doi.org/10.1007/BF02521844 (1972).
31. Yuan, P. et al. Feasibility of ERA5 Integrated Water Vapor Trends for Climate Change Analysis in continental Europe: An Evaluation with GPS (1994–2019) by Considering Statistical Significance. *Remote Sens. Environ.* https://doi.org/10.1016/j.rse.2021.112416 (2021).

32. Mingari, L. et al. Data assimilation of volcanic aerosol observations using FALL3D+PDAF. *Atmos. Chem. Phys.* **21**, 1773–1792. https://doi.org/10.5194/acp-22-1773-2022 (2022).
33. Goodliff, M. et al. Temperature assimilation into a coastal ocean-biogeochemical model: Assessment of weakly- and strongly-coupled data assimilation. *Ocean Dynamics* **69**, 1217–1237 (2019).
34. Tang, Q., Mu, L., Goessling, H. F., Semmler, T. & Nerger, L. Strongly coupled data assimilation of ocean observations into an ocean-atmosphere model. *Geophys. Res. Lett.* https://doi.org/10.1029/2021GL094941 (2021).
35. Zheng, Y., Albergel, C., Munier, S., Bonan, B. & Calvet, J.-C. An offline framework for high-dimensional ensemble Kalman filters to reduce the time to solution. *Geosci. Model Dev.* **13**, 3607–3625. https://doi.org/10.5194/gmd-13-3607-2020 (2020).
36. Li, Y., Cong, Z. & Yang, D. (2023) remotely sensed soil moisture assimilation in the distributed hydrological model based on the error subspace transform Kalman filter. *Remote Sens.* **15**, 7. https://doi.org/10.3390/rs15071852 (1852).
37. Pham, D. T., Verron, J. & Roubaud, M. C. A singular evolutive extended Kalman filter for data assimilation in oceanography. *J. Mar. Syst.* **16**, 323–340. https://doi.org/10.1016/S0924-7963(97)00109-7 (1998).
38. Gaspari, G. & Cohn, S. E. Construction of correlation functions in two and three dimensions. *Q. J. R. Meteorol. Soc.* **125**, 723–757 (1999).
39. Gropp, W., Lusk, E. & Skjellum, A. *Using MPI: Portable Parallel Programming with the Message-Passing Interface* (The MIT Press, 1994).
40. OpenMP. (2008) OpenMP Application Program Interface Version 3.0, http://www.openmp.org/ (last access: 26 June 2023).
41. Bao, X. & Zhang, F. Evaluation of NCEP–CFSR, NCEP–NCAR, ERA-Interim, and ERA-40 Reanalysis Datasets against Independent Sounding Observations over the Tibetan Plateau. *J. Clim.* **26**, 206–214. https://doi.org/10.1175/JCLI-D-12-00056.1 (2013).
42. Pawel, H., Jaroslaw, B. & Witold, R. Assessment of errors in Precipitable Water data derived from Global Navigation Satellite System observations. *J. Atmos. Solar-Terrestrial Physics* **129**, 69–77. https://doi.org/10.1016/j.jastp.2015.04.012 (2015).
43. Li, L., Žagar, N., Raeder, K. & Anderson, J. L. Comparison of temperature and wind observations in the Tropics in a perfect-model, global EnKF data assimilation system. *Quarterly Journal of The Royal Meteorological Society* **149**, 2376–2385. https://doi.org/10.1002/qj.4511 (2023).
44. Van Den Dool, H. M. A new look at weather forecasting through analogues. *Monthly Weather Rev.* **117**(10), 2230–2247 (1989).
45. Talagrand, O., Vautard, R. and Strauss, B. Evaluation of probabilistic prediction systems, in Workshopon Predictability, 20–22 October 1997,1–26, ECMWF, Shinfield Park, Reading. (1997).
46. Rohm, W., Guzikowski, J., Wilgan, K. & Kryza, M. 4DVAR assimilation of GNSS zenith path delays and precipitable water into a numerical weather prediction model WRF. *Atmos. Measur. Tech.* **12**, 345–361. https://doi.org/10.5194/amt-12-345-2019 (2019).
47. Bai, W. *et al.* Applications of GNSS-RO to numerical weather prediction and tropical cyclone forecast. *Atmosphere* **11**, 1204. https://doi.org/10.3390/atmos11111204 (2020).
48. Yang, S. C., Chen, S. H. & Chang, C. C. Understanding the impact of assimilating FORMOSAT-7/COSMIC-2 radio occultation refractivity on tropical cyclone genesis: Observing system simulation experiments using Hurricane Gordon (2006) as a case study. *Q. J. R. Meteorol. Soci.* **149**(753), 1293–1318 (2023).
49. Whitaker, J. S., Hamill, T. M., Wei, X., Song, Y. & Toth, Z. Ensemble data assimilation with the NCEP Global Forecast System. *Monthly. Weather Rev.* **136**, 463–481. https://doi.org/10.1175/2007MWR2018.1 (2008).
50. Sobash, R. A. & Stensrud, D. J. The impact of covariance localization for radar data on EnKF analyses of a developing MCS: Observing system simulation experiments. *Monthly Weather Rev.* **141**, 3691–3709. https://doi.org/10.1175/MWR-D-12-00203.1 (2013).
51. Periáñez, Á., Reich, H. & Potthast, R. Optimal localization for ensemble Kalman filter systems. *J. Meteorol. Soc. Japan* **92**, 585–597. https://doi.org/10.2151/jmsj.2014-605 (2014).
52. Kirchgessner, P., Nerger, L. & Bunse-Gerstner, A. On the choice of an optimal localization radius in ensemble kalman filter methods. *Monthly Weather Rev.* **142**(6), 2165–2175. https://doi.org/10.1175/MWR-D-13-00246.1 (2014).
53. Campagnolo, M. L. et al. Estimating the effective spatial resolution of the operational BRDF, albedo, and nadir reflectance products from MODIS and VIIRS. *Remote Sens. Environ.* **175**, 52–64. https://doi.org/10.1016/j.rse.2015.12.033 (2016).
54. Klaver, R., Haarsma, R., Vidale, P. L. & Hazeleger, W. Effective resolution in high resolution global atmospheric models for climate studies. *Atmos. Sci. Lett.* https://doi.org/10.1002/asl.952 (2020).
55. Hsu, C.-T., Matsuo, T. & Liu, J.-Y. Impact of assimilating the FORMOSAT-3/COSMIC and FORMOSAT-7/COSMIC-2 RO data on the midlatitude and low-latitude ionospheric specification. *Earth Space Sci.* **5**, 875–890. https://doi.org/10.1029/2018EA000447 (2018).
56. Leidner, S. M. et al. A severe weather quick observing system simulation experiment (QuickOSSE) of global navigation satellite system (GNSS) radio occultation (RO) superconstellations. *Monthly Weather Rev.* **145**(2), 637–651. https://doi.org/10.1175/MWR-D-16-0212.1 (2017).
57. Mueller, M. J. et al. Impact of refractivity profiles from a Proposed GNSS-RO constellation on tropical cyclone forecasts in a global Modeling system. *Monthly Weather Rev.* **148**(7), 3037–3057. https://doi.org/10.1175/MWR-D-19-0360.1 (2020).
58. Privé, N. C., McGrath-Spangler, E. L., Carvalho, D., Karpowicz, B. M. & Moradi, I. Robustness of observing system simulation experiments. *Tellus A Dynamic Meteorol. Oceanogr.* **75**(1), 309–333. https://doi.org/10.16993/tellusa.3254 (2023).
59. Wang, L. et al. Orbit Simulator for Satellite and Near-Space Platforms Supporting Observing System Simulation Experiments. *J. Atmos. Oceanic Technol.* **38**(12), 2109–2123. https://doi.org/10.1175/JTECH-D-21-0066.1 (2021).
60. Xie, J., Bertino, L., Cardellach, E., Semmling, M. & Wickert, J. An osse evaluation of the gnss-r altimetry data for the geros-iss mission as a complement to the existing observational networks. *Remote Sens. Environ.* **209**, 152–165. https://doi.org/10.1016/j.rse.2018.02.053 (2018).

Acknowledgements

The calculations for this research were conducted on the high-performance computer of the Alfred Wagner Institute.

Author contributions

Changliang Shao and Lars Nerger planned the campaign; Changliang Shao performed the experiments, analysed the data and wrote the manuscript draft; Lars Nerger reviewed and edited the manuscript.

Funding

Changliang Shao (No. 202105330044) is supported by the China Scholarship Council for one-year research at AWI, the Joint Open Project of KLME & CIC-FEMD, NUIST (KLME202407) and Key Laboratory of Space Ocean Remote Sensing and Application, Ministry of Natural Resources (202402001).

Declaration

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to C.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.