# Data Assimilation in High Dimensional Systems Using Local Particle Filters
## Overcoming the curse of dimensionality in hydrology

Wang, Z.

**DOI**
10.4233/uuid:a7c34b83-8e01-4c54-a27b-bb202500abfd

**Publication date**
2021

**Document Version**
Final published version

**Citation (APA)**
Wang, Z. (2021). *Data Assimilation in High Dimensional Systems Using Local Particle Filters: Overcoming the curse of dimensionality in hydrology*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:a7c34b83-8e01-4c54-a27b-bb202500abfd

# Data Assimilation in High Dimensional Systems Using Local Particle Filters

Overcoming the curse of dimensionality in hydrology

# Data Assimilation in High Dimensional Systems Using Local Particle Filters

## Overcoming the curse of dimensionality in hydrology

**Dissertation**

for the purpose of obtaining the degree of doctor
at Delft University of Technology,
by the authority of the Rector Magnificus Prof.dr.ir. T.H.J.J. van der Hagen,
chair of the Board for Doctorates,
to be defended publicly on Monday, 21 June, 2021 at 12:30 am

by

**Zhenwu WANG**

Master of Engineering in Agricultural Water-Soil Engineering
Zhejiang University, China,
born in Xinzhou, Shanxi, China.

This dissertation has been approved by the promotors.

Promotor: Prof.dr.ir. N.C. van de Giesen
Copromotor: Dr.ir. R.W. Hut

Composition of the doctoral committee:

| | |
|---|---|
| Rector Magnificus, | chairperson |
| Prof.dr.ir. N.C. van de Giesen, | Delft University of Technology, promotor |
| Dr.ir. R.W. Hut, | Delft University of Technology, copromotor |
| | |
| *Independent members:* | |
| Dr.ir. N. Wanders, | Utrecht University |
| Dr. J. Dong, | Massachusetts Institute of Technology |
| Prof.dr.ir. S.C. Steele Dunne, | Delft University of Technology |
| Prof.dr.ir. G. De Lannoy, | Katholieke Universiteit Leuven |
| Prof.dr.ir. A.W. Heemink, | Delft University of Technology |

This research is funded by Delft University of Technology and the Chinese Scholarship Council (CSC) with the project reference number of 201606320249.

*Keywords:* Data assimilation, Particle filters, Localization, High-dimensional systems, Satellite soil moisture, PCR-GLOBWB 2.0 model, Discharge estimation

*Printed by:* Ridderprint | www.ridderprint.nl.

*Front & Back:* Designed by Zhenwu WANG

ISBN 978-94-6384-228-0

An electronic version of this dissertation is available at
`http://repository.tudelft.nl/`.

To my beloved family

# Summary

This dissertation's ultimate goal is to provide solutions to two problems that the promising data assimilation method, called the Particle Filter, has when applied to high dimensional non-linear models, such as those often used in hydrological research and forecasting. Two local particle filters have been proposed to overcome three major issues. Firstly, the curse of dimensionality caused by high dimensional models. Secondly, the uncertainty brought by the data assimilation method itself and finally the problem of nonlinearity in observation operators that link model states to observations. Both newly introduced data assimilation algorithms have been assessed using the Lorenz model (1996), a toy model that provides a perfect evaluation environment for such methods because it is a one-dimensional discrete chaotic model, which can simulate the behavior of changes of atmosphere. One local particle filter has been used in a practical application in hydrology to improve discharge accuracy in the Rhine river basin by assimilating satellite soil moisture into the PCR-GLOWB hydrological model.

The curse of dimensionality is well-known in particle filters. It happens in high dimensional models because, to remain accurate, the number of particles needs to increase exponentially with the increase of the model scale (ie. model dimension). One possible solution to avoid this curse is to apply localization in particle filters. Both proposed particle filters are based on a localization method. Uncertainty sources in data assimilation are many, and it is not easy to separate all of them clearly and directly. The two variants of the particle filter proposed in this thesis focus on different issues.

The localization used in the first particle filters divided the whole analysis of data assimilation into small batches for each model state. Each local analysis is independent, and it only assimilates observations within the localization scale. In the process it quantifies the uncertainty that is introduced by the data assimilation process itself. The localization method for the second local particle filter variant used another strategy. In its procedure, all observations are assimilated one by one, and each observation only affects near model states within the localization radius. When all observations are assimilated sequentially, all model states are updated. In addition, the second particle filter variant tried to solve the problem caused by nonlinear observation operators. To overcome the latter problems, the nonlinear observation operator was replaced by a surrogate model, named the Gaussian process regression model. For the calculation of the weights for each particle, model states needed to be transferred into the observation space. A Gaussian process regression surrogate model makes the transition process more straightforward in the nonlinear case because it provides the mean and standard deviation of estimates. Both local particle filter variants introduced in this thesis were evaluated thoroughly, and all results demonstrated that they performed satisfactorily in the specific nonlinear case and can be applied in high dimensional systems.

In addition to testing both local particle filters in the controlled Lorenz model, LPF-GT has also been verified as beneficial in a case study with the hydrological model PCR-

GLOBWB. The specific study area focused on the Rhine river basin. The local particle filters have been applied to assimilate satellite soil moisture from the SMAP mission into the PCR-GLOBWB model to improve discharge estimates. Results show that the local particle filter performed well and significantly improved discharge accuracy by assimilating SMAP soil moisture. The new LPF-GT only requires a handful of particles to reach better performance in the Rhine river basin. This is particularly useful and practical for large-scale models that are often used in hydrology. Only requiring a small number of particles is the primary advantage of this data assimilation method because it saves lots of computational costs. In addition, the use of the localization in this particle filter makes the update for each model state independent from each other and can be conducted in parallel. Thus, the efficiency of this data assimilation method can be improved further.

In conclusion, the new additions to the particle filter proposed in this thesis are stable and can provide satisfying accuracy in nonlinear cases and for high dimensional models. Both of them have been proven to perform well in a toy model with many dimensions where they have direct value in solving the curse of dimensionality and nonlinearity. More importantly, they are valuable data assimilation methods to give direct insights into how to cope with uncertainty in nonlinear cases and to offer data assimilation frameworks for developing new particle filters in the future. The successful hydrological application of data assimilation using local particle filters in this research shows its considerable potential in hydrology.

# Contents

# 1

# Introduction

*Those who fail in everyday affairs show a tendency to reach out for the impossible. For when we fall in attempting the possible, the blame is solely ours; but when we fail in attempting the impossible, we are justified in attributing it to the magnitude of the task.*

Eric Hoffer

## 1.1. Data assimilation

Data assimilation (DA) is a field of science that aims to develop algorithms that optimally estimate the continuously changing state of a dynamical system by using all relevant information known about the system combined with, often real-time, observations of the physical system represented by the model. A widely known application of DA is updating yesterdays forecast of today's weather with new observations measured today to arrive at the best estimation of the current state of the weather. This is subsequently used as a starting point for today's estimation of tomorrows weather. Since both yesterday's forecast of today's weather and today's observations will have uncertainties disregarding one, but not the other, is sub-optimal. Scientists in the field of data assimilation develop algorithms that calculate the optimal estimation of today's weather, given all uncertainties involved.

Generally, doing data assimilation needs three essential components: a numerical model representing the system of interest propagated over time, observations collected at different times and places that relate to the state of the system, and a data assimilation algorithm. In general, a model consists of mathematical equations that represent the dominant physical processes in the system. Because we do not have the perfect knowledge of the physical world, it is not easy to define the mathematical model accurately. Therefore, a numerical model has errors, and the uncertainty of a model's estimates usually get larger when the model projects further into the future. Adding information from observations through a data assimilation algorithm can keep the model stable with a relatively satisfying result in the long term. The application of DA is widespread, and it has been used massively in meteorology, ocean science, etc [1–4].

## 1.2. Data assimilation in Hydrology

While applications of DA within the geosciences are best known in atmospheric science, mainly operational weather forecasting, they also see extensive use in operational hydrological forecasting. DA algorithms have been used to enhance the accuracy of hydrological models. For different objectives, various components in hydrology can be improved by assimilating available observations. Several typical DA applications in hydrology are listed below.

1. DA has been used in rainfall-runoff models to assimilate streamflow data in operational flow forecasting systems to obtain improved flow forecasts and predictions of floods [5–8].

2. Numerous studies have focused on the assimilation of surface soil moisture into hydrological models. Soil moisture is a crucial part of hydrology, as its value can switch a model's behavior from slow (groundwater flow) to fast (overland flow) and is thus very important in flood forecasting. Improving its estimations with data assimilation of observations definitely improves the predictive power of hydrological models [9–15].

3. The impact of evaporation DA on hydrological processes has been investigated mainly because terrestrial actual evaporation is an import component of the terrestrial watercycle. Evaporation DA provided improved regional evaporation es-

timates. Better model predictions of soil moisture and streamflow are achieved by assimilating evaporation into the hydrological model [16–21].

4. Assimilating water level data derived from satellite products to estimate discharge is feasible. The DA of water level data has great potential to reduce discharge uncertainty [22–26].

5. Snow, as an essential part of the hydrological cycle, plays a vital role in Earth's energy balance. Assimilation of snow information into models in Earth sciences is important to address the impact of snow on the hydrological and weather forecast to predict snow-related water resources [27–32].

6. Leaf area index (LAI) is a critical environmental variable, providing feedback on vegetation for hydrological, land surface, and climate models [33]. Assimilation of LAI could improve the accuracy of soil moisture [34] and water fluxes [35–38].

Thanks to recent satellite developments for hydrology, currently, satellite data products are the primary sources of observations in hydrological DA applications. Assimilating satellite soil moisture [39] or GRACE data [40–48] into dynamic hydrological models leads to improved estimation of multiple components, or states, of the water system.

## 1.3. Particle filters, non-Gaussian filters

The Ensemble Kalman Filter [49, EnKF] and its variants are popular and commonly used in hydrology and many other Earth science fields. Ensemble-type filters are based on Monte Carlo methods, and its analysis step relies on a Gaussian assumption, which is its main limitation. In nonlinear and non-Gaussian systems, ensemble-type filters are sub-optimal and provide poor estimates of model states. Unfortunately, most observations do have a non-Gaussian error distribution in Earth science. Most geophysical systems are nonlinear, and consequently, model errors are non-Gaussian after the process of model propagation [50]. Moreover, other sources of non-linearity and non-Gaussianity in modeling and observations, such as thresholds of microphysics and higher model resolution that need better physical simulation [50], can lead to the collapse of EnKF.

Compared with ensemble-type filters, a particle filter [50–55] relaxes all linear and Gaussian assumptions, and allows a full Bayesian analysis, which are appealing properties of the particle filter and making it particularly promising.

## 1.4. Filter collapse

Particle filters have been applied successfully in low-dimensional models [56]. But for cases with higher-dimensional models, PFs inevitable suffer from weight degeneracy [55, 57, 58]. When the number of dimensions of a model is low, the weights of particles are balanced, and the variance of particles has a value not close to zero. However, as the model dimension increases, weight degeneracy happens quickly, and all weights have the same value. Consequently, the variance of particles becomes zero. This phenomenon is well-known in the PF literature and is generally called the curse of dimensionality, filter collapse, filter degeneracy, or filter impoverishment.

Bocquet *et al.* [50] used the Lorenz-96 model (1996) to show weight degeneracy in PFs using 128 particles. When the model dimension is smaller than 20, PFs can work stably, and the variance of particles keeps at a particular value. Weights degenerate rapidly when the model has 40 variables. Farchi and Bocquet [58] demonstrated the curse of dimensionality by using a Gaussian linear model. In this case, weight degeneracy happens when the model dimension grows to 32. Using more particles can avoid filter collapse. But the required number of particles to prevent filter collapse scales exponentially with the dimension of a system [60, 61]. In real applications of Earth sciences, the model typically has hundreds and thousands of variables, and the need for particles is substantial. Consequently, a huge amount of memory is needed to store all those particles, which is prohibitive and impossible in practice.

## 1.5. Localization, a way to avoid filter collapse

Localization has been commonly used in Ensemble-type filters and has been proven to be effective [62]. The basic idea of localization is to update model state variables by assimilating observations within a particular (local, regional) scale. Distant observations are excluded when the distance is too far away from variables. The reasoning behind localisation is that there is a maximum distance over which observations can still influence a state, ie: soil moisture measurements in one part of the world are unlikely to effect (or: be spatially correlated with) the soil moisture state in another part of the world.

Generally, there are two ways to apply localization in ensemble-type filters. Either by operating on background error covariances (called B-localization) or observation error covariances (called R-localization). Because the update process of PFs does not rely on the error covariance, the implementation of localization in PFs is slightly different from that in EnKfs. We can divide the solutions into two categories based on where the local analysis is performed. Suppose all model state variables are updated independently by only using observations centered on each grid point within a certain radius. In that case, we can call it state-domain localization [58] and it has been used in several studies [63–66]. When observations are assimilated sequentially, only nearby grid points are updated and influenced, and we can call it sequential-observation localization [58], which was proposed by Poterjoy [67], and has been applied in a simplified atmospheric model [68]. Localization in PFs can beat the curse of dimensionality successfully. For state-domain localization, it is easy to parallelize, but this update scheme may cause the consistency issue because the relationship between state variables is broken. The implementation of sequential-observation localization is harder to parallelize but still possible, but it may alleviate the consistency issue [58].

## 1.6. Knowledge gap

As mentioned previously, localization methods provide an effective solution to the curse of dimensionality in particle filters and make it possible to implement PFs in high dimensional models. Based on localization, which is an essential foundation of this research, other particle filters' challenges become the central focus of this research.

Uncertainty can arise from various sources in a data assimilation cycle [62]. For example, sample errors due to the limited number of particles or ensemble members

[62], incorrect observation errors [69–71], and multi parameterization [72]. The list of additional error sources can be extended further. However, in practice, it is difficult to distinguish and account for all error sources. Currently, we do not have a proper way to quantify and describe model errors. It is unclear how model uncertainty can be quantified to improve the performance of particle filters. A better understanding of errors in estimation of the state while using a DA algorithm would provide a significant improvement.

Nonlinear issues in data assimilation have attracted considerable attention recently [73]. Increasing computational power makes it possible to run operational models with many dimensions. This approach could bring more nonlinearity because of the existence of small-scale nonlinear processes in these models. Apart from this, observation networks worldwide provide more and more new products with higher accuracy and higher resolution. The nonlinear observation operators, which links model states and observations, requires non-Gaussian data assimilation methods [58, 66, 74–78].

For nonlinear and high dimensional systems, exploring and developing nonlinear filters beyond Gaussian assumptions improves estimations of system states using data assimilation. Since particle filters are not limited by Gaussian assumptions, proposing new particle filters, or hybrid filters between particle filters and ensemble-based methods, provides the potential to cope with nonlinear issues in data assimilation.

To meet these challenges, appropriate data assimilation strategies need to be developed to bridge the knowledge gap. This thesis aims to develop particle filters with localization to capture and quantify possible sources of uncertainty in particle filters and to overcome problems with nonlinear observation operators.

## 1.7. Research outline

The rest of this dissertation is organized as follows:

Chapter 2 describes several fundamental theories used in the next two chapters, including particle filters, Gamma test theory, and Gaussian process regression.

Chapter 3 introduces a novel variant of the local particle filter with localization B, the Gamma test theory. The description of this filter is given in detail. A set of experiments was conducted to evaluate the proposed method's performance by using a Lorenz model. This new filter considers the uncertainty brought by the process of data assimilation and is applied in a high-dimensional system.

Chapter 4 presents another variant of a local particle filter using Gaussian process regression models with the C localization method. The structure of this chapter is similar to the last chapter. The algorithm was further elaborated, and several experiments with various configurations were performed to test its performance. Typically, it is common to use the Lorenz model to evaluate a new data assimilation method. Therefore, all experiments are based on it. This new filter's primary goal is to solve the issues caused by nonlinear observation operators in a high-dimensional model.

Chapter 5 builds on the previous chapters and applies the newly introduced particle filter with localization mentioned in chapter 3 to the hydrological model PCR-GLOBWB. SMAP soil moisture data are assimilated into the PCR-GLOBWB model to improve the estimation of discharge in the Rhine basin. Considering the advantages of particle filters over ensemble-type methods, which is given in section 1.3, it is worth exploring particle filters' possibilities in a real hydrological application.

Chapter 6 summarizes the contributions of this thesis in particle filters with localization. Limitations of this research and potential future research on particle filters and data assimilation in hydrology are listed and discussed in a broader context.

## References

[1] M. A. Balmaseda, D. Dee, A. Vidard, and D. L. T. Anderson, *A multivariate treatment of bias for sequential data assimilation: Application to the tropical oceans,* Quarterly Journal of the Royal Meteorological Society 133, 167 (2007).

[2] N. Bormann and P. Bauer, *Estimates of spatial and interchannel observation-error characteristics for current sounder radiances for numerical weather prediction. I: Methods and application to ATOVS data,* Quarterly Journal of the Royal Meteorological Society 136, 1036 (2010).

[3] I. Hoteit, X. Luo, and D.-T. Pham, *Particle Kalman Filtering: A Nonlinear Bayesian Framework for Ensemble Kalman Filters,* Monthly Weather Review 140, 528 (2012).

[4] C. A. Edwards, A. M. Moore, I. Hoteit, and B. D. Cornuelle, *Regional Ocean Data Assimilation,* Annual Review of Marine Science 7, 21 (2015).

[5] P. Abbaszadeh, K. Gavahi, and H. Moradkhani, *Multivariate remotely sensed and in-situ data assimilation for enhancing community WRF-Hydro model forecasting,* Advances in Water Resources 145, 103721 (2020).

[6] M. Ramgraber, M. Camporese, P. Renard, P. Salandin, and M. Schirmer, *Quasi-Online Groundwater Model Optimization Under Constraints of Geological Consistency Based on Iterative Importance Sampling,* Water Resources Research 56, e2019WR026777 (2020).

[7] D. H. Kim, *High-spatial-resolution streamflow estimation at ungauged river sites or gauged sites with missing data using the National Hydrography Dataset (NHD) and U.S. Geological Survey (USGS) streamflow data,* Journal of Hydrology 565, 819 (2018).

[8] G. Ercolani and F. Castelli, *Variational assimilation of streamflow data in distributed flood forecasting,* Water Resources Research 53, 158 (2017).

[9] M. Ciupak, B. Ozga-Zielinski, J. Adamowski, R. C. Deo, and K. Kochanek, *Correcting Satellite Precipitation Data and Assimilating Satellite-Derived Soil Moisture Data to Generate Ensemble Hydrological Forecasts within the HBV Rainfall-Runoff Model,* Water 11, 2138 (2019).

[10] N. Raoult, C. Ottlé, P. Peylin, V. Bastrikov, and P. Maugis, *Evaluating and Optimizing Surface Soil Moisture Drydowns in the ORCHIDEE Land Surface Model at In Situ Locations,* Journal of Hydrometeorology 22, 1025 (2021).

[11] E. Pinnington, J. Amezcua, E. Cooper, S. Dadson, R. Ellis, J. Peng, E. Robinson, R. Morrison, S. Osborne, and T. Quaife, *Improving soil moisture prediction of*

*a high-resolution land surface model by parameterising pedotransfer functions through assimilation of SMAP satellite data,* Hydrology and Earth System Sciences 25, 1617 (2021).

[12] P. Baguis and E. Roulin, *Soil Moisture Data Assimilation in a Hydrological Model: A Case Study in Belgium Using Large-Scale Satellite Data,* Remote Sensing 9, 820 (2017).

[13] J. Eeckman, H. Roux, A. Douinot, B. Bonan, and C. Albergel, *A multi-sourced assessment of the spatiotemporal dynamics of soil moisture in the MARINE flash flood model,* Hydrology and Earth System Sciences 25, 1425 (2021).

[14] N. Jadidoleslam, R. Mantilla, and W. F. Krajewski, *Data Assimilation of Satellite-Based Soil Moisture into a Distributed Hydrological Model for Streamflow Predictions,* Hydrology 8, 52 (2021).

[15] Y. Wang, L. Shi, T. Xu, Q. Zhang, M. Ye, and Y. Zha, *A nonparametric sequential data assimilation scheme for soil moisture flow,* Journal of Hydrology 593, 125865 (2021).

[16] J. M. Schuurmans, P. A. Troch, A. A. Veldhuizen, W. G. M. Bastiaanssen, and M. F. P. Bierkens, *Assimilation of remotely sensed latent heat flux in a distributed hydrological model,* Advances in Water Resources 26, 151 (2003).

[17] E. P. Glenn, C. M. U. Neale, D. J. Hunsaker, and P. L. Nagler, *Vegetation index-based crop coefficients to estimate evapotranspiration by remote sensing in agricultural and natural ecosystems,* Hydrological Processes 25, 4050 (2011).

[18] T. Xu, S. M. Bateni, S. Liang, D. Entekhabi, and K. Mao, *Estimation of surface turbulent heat fluxes via variational assimilation of sequences of land surface temperatures from Geostationary Operational Environmental Satellites,* Journal of Geophysical Research: Atmospheres 119, 10,780 (2014).

[19] L. Zou, C. Zhan, J. Xia, T. Wang, and C. J. Gippel, *Implementation of evapotranspiration data assimilation with catchment scale distributed hydrological model via an ensemble Kalman Filter,* Journal of Hydrology 549, 685 (2017).

[20] F. Pourmansouri and M. Rahimzadegan, *Evaluation of vegetation and evapotranspiration changes in Iran using satellite data and ground measurements,* Journal of Applied Remote Sensing 14, 034530 (2020).

[21] S. Gelsinari, R. Doble, E. Daly, and V. R. N. Pauwels, *Feasibility of Improving Groundwater Modeling by Assimilating Evapotranspiration Rates,* Water Resources Research 56, e2019WR025983 (2020).

[22] J. Neal, G. Schumann, P. Bates, W. Buytaert, P. Matgen, and F. Pappenberger, *A data assimilation approach to discharge estimation from space,* Hydrological Processes 23, 3641 (2009).

[23] J. Santos da Silva, S. Calmant, F. Seyler, O. C. Rotunno Filho, G. Cochonneau, and W. J. Mansur, *Water levels in the Amazon basin derived from the ERS 2 and ENVISAT radar altimetry missions,* Remote Sensing of Environment 114, 2160 (2010).

[24] L. Giustarini, P. Matgen, R. Hostache, M. Montanari, D. Plaza, V. R. N. Pauwels, G. J. M. De Lannoy, R. De Keyser, L. Pfister, L. Hoffmann, and H. H. G. Savenije, *Assimilating SAR-derived water level data into a hydraulic model: A case study,* Hydrology and Earth System Sciences 15, 2349 (2011).

[25] A. Asadzadeh Jarihani, J. N. Callow, K. Johansen, and B. Gouweleeuw, *Evaluation of multiple satellite altimetry data for studying inland water bodies and river floods,* Journal of Hydrology 505, 78 (2013).

[26] M. Khaki, E. Forootan, and M. A. Sharifi, *Satellite radar altimetry waveform retracking over the Caspian Sea,* International Journal of Remote Sensing 35, 6329 (2014).

[27] K. N. Musselman, N. Addor, J. A. Vano, and N. P. Molotch, *Winter melt trends portend widespread declines in snow water resources,* Nature Climate Change 11, 418 (2021).

[28] C. Largeron, M. Dumont, S. Morin, A. A. Boone, M. Lafaysse, S. Metref, E. Cosme, T. Jonas, A. Winstral, and S. A. Margulis, *Toward Snow Cover Estimation in Mountainous Areas Using Modern Data Assimilation Methods: A Review,* Frontiers in Earth Science 8, 325 (2020).

[29] F. Appel, F. Koch, A. Rösel, P. Klug, P. Henkel, M. Lamm, W. Mauser, and H. Bach, *Advances in Snow Hydrology Using a Combined Approach of GNSS In Situ Stations, Hydrological Modelling and Earth Observation—A Case Study in Canada,* Geosciences 9, 44 (2019).

[30] J. Helmert, A. Şensoy Şorman, R. Alvarado Montero, C. De Michele, P. De Rosnay, M. Dumont, D. C. Finger, M. Lange, G. Picard, V. Potopová, S. Pullen, D. Vikhamar-Schuler, and A. N. Arslan, *Review of Snow Data Assimilation Methods for Hydrological, Land Surface, Meteorological and Climate Models: Results from a COST HarmoSnow Survey,* Geosciences 8, 489 (2018).

[31] C. Huang, A. J. Newman, M. P. Clark, A. W. Wood, and X. Zheng, *Evaluation of snow data assimilation using the ensemble Kalman filter for seasonal streamflow prediction in the western United States,* Hydrology and Earth System Sciences 21, 635 (2017).

[32] Y.-F. Zhang and Z.-L. Yang, *Estimating uncertainties in the newly developed multi-source land snow data assimilation system,* Journal of Geophysical Research: Atmospheres 121, 8254 (2016).

[33] H. Fang, F. Baret, S. Plummer, and G. Schaepman-Strub, *An Overview of Global Leaf Area Index (LAI): Methods, Products, Validation, and Applications,* Reviews of Geophysics 57, 739 (2019).

[34] C. Albergel, S. Munier, D. J. Leroux, H. Dewaele, D. Fairbairn, A. L. Barbu, E. Gelati, W. Dorigo, S. Faroux, C. Meurey, P. Le Moigne, B. Decharme, J.-F. Mahfouf, and J.-C. Calvet, *Sequential assimilation of satellite-derived vegetation and soil moisture products using SURFEX_v8.0: LDAS-Monde assessment over the Euro-Mediterranean area,* Geoscientific Model Development 10, 3889 (2017).

[35] C. Albergel, S. Munier, A. Bocher, B. Bonan, Y. Zheng, C. Draper, D. J. Leroux, and J.-C. Calvet, *LDAS-Monde Sequential Assimilation of Satellite Derived Observations Applied to the Contiguous US: An ERA-5 Driven Reanalysis of the Land Surface Variables,* Remote Sensing 10, 1627 (2018).

[36] S. V. Kumar, D. M. Mocko, S. Wang, C. D. Peters-Lidard, and J. Borak, *Assimilation of Remotely Sensed Leaf Area Index into the Noah-MP Land Surface Model: Impacts on Water and Carbon Fluxes and States over the Continental United States,* Journal of Hydrometeorology 20, 1359 (2019).

[37] T. B. Ramos, L. Simionesei, A. R. Oliveira, H. Darouich, and R. Neves, *Assessing the Impact of LAI Data Assimilation on Simulations of the Soil Water Balance and Maize Development Using MOHID-Land,* Water 10, 1367 (2018).

[38] H. Seo and Y. Kim, *Role of remotely sensed leaf area index assimilation in ecohydrologic processes in different ecosystems over East Asia with Community Land Model version 4.5 – Biogeochemistry,* Journal of Hydrology 594, 125957 (2021).

[39] L. J. Renzullo, A. I. J. M. van Dijk, J. M. Perraud, D. Collins, B. Henderson, H. Jin, A. B. Smith, and D. L. McJannet, *Continental satellite soil moisture data assimilation improves root-zone moisture analysis for water resources assessment,* Journal of Hydrology 519, 2747 (2014).

[40] C. Yao, Z. Luo, H. Wang, Q. Li, and H. Zhou, *GRACE-Derived Terrestrial Water Storage Changes in the Inter-Basin Region and Its Possible Influencing Factors: A Case Study of the Sichuan Basin, China,* Remote Sensing 8, 444 (2016).

[41] N. Tangdamrongsub, S. C. Steele-Dunne, B. C. Gunter, P. G. Ditmar, and A. H. Weerts, *Data assimilation of GRACE terrestrial water storage estimates into a regional hydrological model of the Rhine River basin,* Hydrology and Earth System Sciences 19, 2079 (2015).

[42] M. Girotto, R. Reichle, M. Rodell, and V. Maggioni, *Data Assimilation of Terrestrial Water Storage Observations to Estimate Precipitation Fluxes: A Synthetic Experiment,* Remote Sensing 13, 1223 (2021).

[43] M. Schumacher, J. Kusche, and P. Döll, *A systematic impact assessment of GRACE error correlation on data assimilation in hydrological models,* Journal of Geodesy 90, 537 (2016).

[44] E. C. Massoud, Z. Liu, A. Shaban, and M. El Hage, *Groundwater Depletion Signals in the Beqaa Plain, Lebanon: Evidence from GRACE and Sentinel-1 Data,* Remote Sensing 13, 915 (2021).

[45] M. Girotto, G. J. M. De Lannoy, R. H. Reichle, M. Rodell, C. Draper, S. N. Bhanja, and A. Mukherjee, *Benefits and pitfalls of GRACE data assimilation: A case study of terrestrial water storage depletion in India,* Geophysical Research Letters 44, 4107 (2017).

[46] M. Girotto, G. J. M. D. Lannoy, R. H. Reichle, and M. Rodell, *Assimilation of gridded terrestrial water storage observations from GRACE into a land surface model,* Water Resources Research 52, 4164 (2016).

[47] M. Khaki, M. Schumacher, E. Forootan, M. Kuhn, J. L. Awange, and A. I. J. M. van Dijk, *Accounting for spatial correlation errors in the assimilation of GRACE into hydrological models through localization,* Advances in Water Resources 108, 99 (2017).

[48] A. Shokri, J. P. Walker, A. I. J. M. van Dijk, and V. R. N. Pauwels, *On the Use of Adaptive Ensemble Kalman Filtering to Mitigate Error Misspecifications in GRACE Data Assimilation,* Water Resources Research 55, 7622 (2019).

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

[50] M. Bocquet, C. A. Pires, and L. Wu, *Beyond Gaussian Statistical Modeling in Geophysical Data Assimilation,* Monthly Weather Review 138, 2997 (2010).

[51] N. J. Gordon, D. J. Salmond, and A. F. M. Smith, *Novel approach to nonlinear/non-Gaussian Bayesian state estimation,* IEE Proceedings F (Radar and Signal Processing) 140, 107 (1993).

[52] M. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, *A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking,* IEEE Transactions on Signal Processing 50, 174 (2002).

[53] A. Doucet, N. de Freitas, and N. Gordon, *An Introduction to Sequential Monte Carlo Methods,* in *Sequential Monte Carlo Methods in Practice*, Statistics for Engineering and Information Science (Springer, New York, NY, 2001) pp. 3–14.

[54] Z. Chen, *Bayesian filtering: From Kalman filters to particle filters, and beyond,* Statistics 182, 1 (2003).

[55] P. J. van Leeuwen, *Particle Filtering in Geophysical Systems,* Monthly Weather Review 137, 4089 (2009).

[56] A. Doucet, S. Godsill, and C. Andrieu, *On sequential Monte Carlo sampling methods for Bayesian filtering,* Statistics and Computing 10, 197 (2000).

[57] Y. Zhou, D. McLaughlin, and D. Entekhabi, *Assessing the Performance of the Ensemble Kalman Filter for Land Surface Data Assimilation,* Monthly Weather Review 134, 2128 (2006).

[58] A. Farchi and M. Bocquet, *Review article: Comparison of local particle filters and new implementations,* Nonlinear Processes in Geophysics 25, 765 (2018).

[59] E. N. Lorenz, *Predictability: A problem partly solved,* in *Proc. Seminar on Predictability, Reading, Uk, Ecmwf* (1996).

[60] C. Snyder, T. Bengtsson, P. Bickel, and J. Anderson, *Obstacles to High-Dimensional Particle Filtering,* Monthly Weather Review 136, 4629 (2008).

[61] L. Slivinski and C. Snyder, *Exploring Practical Estimates of the Ensemble Size Necessary for Particle Filters,* Monthly Weather Review 144, 861 (2016).

[62] P. L. Houtekamer and F. Zhang, *Review of the Ensemble Kalman Filter for Atmospheric Data Assimilation,* Monthly Weather Review 144, 4489 (2016).

[63] S. G. Penny and T. Miyoshi, *A local particle filter for high dimensional geophysical systems,* Nonlinear Processes in Geophysics Discussions 2, 1631 (2015).

[64] Y. Lee and A. J. Majda, *State estimation and prediction using clustered particle filters,* Proceedings of the National Academy of Sciences 113, 14609 (2016).

[65] P. Rebeschini and R. van Handel, *Can local particle filters beat the curse of dimensionality?* The Annals of Applied Probability 25, 2809 (2015).

[66] N. Chustagulprom, S. Reich, and M. Reinhardt, *A Hybrid Ensemble Transform Particle Filter for Nonlinear and Spatially Extended Dynamical Systems,* SIAM/ASA Journal on Uncertainty Quantification 4, 592 (2016).

[67] J. Poterjoy, *A Localized Particle Filter for High-Dimensional Nonlinear Systems,* Monthly Weather Review 144, 59 (2016).

[68] J. Poterjoy and J. L. Anderson, *Efficient Assimilation of Simulated Observations in a High-Dimensional Geophysical System Using a Localized Particle Filter,* Monthly Weather Review 144, 2007 (2016).

[69] R. Frehlich, *Adaptive data assimilation including the effect of spatial variations in observation error,* Quarterly Journal of the Royal Meteorological Society 132, 1225 (2006).

[70] V. E. Gorin and M. D. Tsyrulnikov, *Estimation of Multivariate Observation-Error Statistics for AMSU-A Data,* Monthly Weather Review 139, 3765 (2011).

[71] L. M. Stewart, S. L. Dance, and N. K. Nichols, *Data assimilation with correlated observation errors: Experiments with a 1-D shallow water model,* Tellus A: Dynamic Meteorology and Oceanography 65, 19546 (2013).

[72] P. L. Houtekamer, H. L. Mitchell, and X. Deng, *Model Error Representation in an Operational Ensemble Kalman Filter,* Monthly Weather Review 137, 2126 (2009).

[73] S. Vetra-Carvalho, P. J. van Leeuwen, L. Nerger, A. Barth, M. U. Altaf, P. Brasseur, P. Kirchgessner, and J.-M. Beckers, *State-of-the-art stochastic data assimilation methods for high-dimensional non-Gaussian problems,* Tellus A: Dynamic Meteorology and Oceanography 70, 1 (2018).

[74] S. Robert and H. R. Künsch, *Localizing the Ensemble Kalman Particle Filter,* Tellus A: Dynamic Meteorology and Oceanography 69, 1282016 (2017).

[75] A. Beskos, D. Crisan, A. Jasra, K. Kamatani, and Y. Zhou, *A stable particle filter for a class of high-dimensional state-space models,* Advances in Applied Probability 49, 24 (2017).

[76] L. Slivinski, E. Spiller, A. Apte, and B. Sandstede, *A Hybrid Particle–Ensemble Kalman Filter for Lagrangian Data Assimilation,* Monthly Weather Review 143, 195 (2015).

[77] P. J. van Leeuwen, *Nonlinear data assimilation in geosciences: An extremely efficient particle filter,* Quarterly Journal of the Royal Meteorological Society 136, 1991 (2010).

[78] R. Potthast, A. Walter, and A. Rhodin, *A Localized Adaptive Particle Filter within an Operational NWP Framework,* Monthly Weather Review 147, 345 (2018).

# 2

# Common ground

*From a mathematical perspective, data assimilation is based on Bayes Theorem. We assume the uncertainty of model states and observations can be represented by a probability measure. Under this assumption, model states and observations can be expressed by a probability density function (PDF). Bayes Theorem updates the PDF of states given observations. Typically, the Monte Carlo method is used to approximate a PDF. It means a PDF is represented by several points, which is sampled from the PDF. In all DA methods, ensemble Kalman filters (EnKF) are the most popular algorithms that are applied widely in Earth sciences. Plenty of new variants of EnKF have been developed to overcome its disadvantages. Particle filters, relaxing the Gaussian assumptions required by ensemble-based filters, have drawn so much attention. The new particle filters, which can defeat the curse of dimensionality, have significant development. All new ideas and fresh insight, possibly leading to new filters, are derived from the basic filters. Detailed descriptions of new local particle filters are given in the next two chapters. They have the same theoretical foundation - particle filters, and the local ensemble transform Kalman filter (LETKF) is chosen as the benchmark to evaluate their performance. To avoid duplication and a lengthy thesis, in this Chapter, we introduced several filters and related theories needed in the following chapters briefly, including standard particle filters, EnKF, LETKF, Gamma test theory (GT), and Gaussian process regression (GPR).*

## 2.1. Standard particle filters

This subsection briefly introduces standard particle filters [1]. Particles are used to represent the distribution of model states, which can capture the mean and uncertainty of the model states. The particles are updated by a resampling algorithm based on weights that are calculated by the likelihood, given observations. The resampling method is crucial to particle filters [2] and its basic idea is to modify prior particles to posterior ones by eliminating particles with smaller weights and by duplicating particles having larger weights. The residual resampling algorithm [3] was applied in this study, as being one of the most frequently used resampling algorithm in particle filter data assimilation [2, 4, 5].

Let us assume that $\boldsymbol{x}$ represents a model state of a model which can be propagated over time and $\boldsymbol{y}$ represents a vector of observations. The model captures our incomplete knowledge of the physical system under consideration. Uncertainty in model states is inevitable due to imperfection of the model. Therefore, the model states, $\boldsymbol{x}$, can only approximate the truth and cannot reach the truth. For similar reasons, observations $\boldsymbol{y}$ are only approximations of the truth because of observation uncertainties. The relation between observations and true model states can be expressed as:

$$\boldsymbol{y} = \boldsymbol{H}\left(\boldsymbol{x}_{\text{true}}\right) + \epsilon \tag{2.1}$$

where $\mathrm{H}$ is the observation operator that transforms model states into observation space and $\epsilon$ is the observation error. Particle filters are used to estimate $\boldsymbol{x}$ given observations $\boldsymbol{y}$ by a Bayes' theorem expansion using a Monte Carlo estimation:

$$p(\boldsymbol{x}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})}{\int p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x}} \tag{2.2}$$

where $p\left(\boldsymbol{x} \mid \boldsymbol{y}\right)$ is the probability of model states $\boldsymbol{x}$ given all observations $\boldsymbol{y}$. $p(\boldsymbol{x}|\boldsymbol{y})$ can be obtained by a Monte Carlo approach. By drawing $N_p$ particles, denoted $\boldsymbol{x}_n(n = 1, 2, 3, ..., N_p)$, $p(\boldsymbol{x})$ can be constructed as a discrete set of delta functions centered on every individual particle:

$$p(\boldsymbol{x}) \approx \frac{1}{N_p}\sum_{n=1}^{N_p}\delta\left(\boldsymbol{x} - \boldsymbol{x}_n\right) \tag{2.3}$$

Similarly, $p(\boldsymbol{x}|\boldsymbol{y})$ can be approximated by using:

$$p(\boldsymbol{x}|\boldsymbol{y}) \approx \sum_{n=1}^{N_p}\frac{\boldsymbol{w_n}}{W}\delta\left(\boldsymbol{x} - \boldsymbol{x}_n\right) \tag{2.4}$$

where $w_n/W$ are the normalized weights, provided by:

$$\boldsymbol{w_n} = p\left(\boldsymbol{y}|\boldsymbol{x}_n\right) \tag{2.5}$$

$$W = \sum_{n=1}^{N_p}\boldsymbol{w_n} \tag{2.6}$$

Generally, a resampling algorithm is used to generate the posterior distribution by discarding particles with low weight and duplicating high weight particles [3]. The resampling process guarantees that sufficient particles remain.

## 2.2. Ensemble Kalman Filter(EnKF)

The EnKF [6] is a variant of the Kalman Filter, and is used as a benchmark in this research. In the EnKF, random samples of model states (together called an ensemble) are used to represent the distributions of model states and observations, including their uncertainty. These ensembles can be generated by adding random errors to model states and observations. Both model and observation error covariances are estimated from the ensemble members. These two error covariance matrices are the fundamental basis of the EnKF. Thus, the success of using the EnKF for data assimilation heavily relies on the accurate estimation of these error characteristics.

The EnKF has two main processing steps, a forward step and an update step, performed sequentially when observations become available. In the forward step, the ensemble of state variables is propagated forward in time. The update step is used to adjust state variables based on the error covariance of model states and observations. The description of the EnKF processing steps is given below. In the forward step, the ensemble of state variables is propagated forward in time as:

$$\boldsymbol{x}_{t+1} = \mathrm{M}\left(\boldsymbol{x}_t, \boldsymbol{u}_t, \boldsymbol{\theta}\right) \tag{2.7}$$

where $\boldsymbol{x}$ is a model state, $M$ is the model operator that propagates state variables over time. $\boldsymbol{u}$ indicates the model forcing, $\boldsymbol{\theta}$ is the model parameters. The model states are related to observations as:

$$\boldsymbol{y} = \boldsymbol{H}\left(\boldsymbol{x}_{\text{true}}\right) + \boldsymbol{\epsilon} \tag{2.8}$$

where $\boldsymbol{y}$ is an observation vector, $\boldsymbol{H}$ is the observation operator that relates model states to observations, and $\boldsymbol{\epsilon}$ is the observation error. The observation error $\boldsymbol{\epsilon}$ is often assumed to follow a Gaussian distribution with zero mean and observation covariance matrix $\boldsymbol{R}$. EnKF is based on the best linear unbiased estimator (BLUE) and the general form of the analysis step in EnKF can be expressed as:

$$\boldsymbol{x}^a = \boldsymbol{x}^f + \boldsymbol{K}\left(\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x}^f\right) \tag{2.9}$$

$$\boldsymbol{K} = \boldsymbol{P}^f\boldsymbol{H}^T\left[\boldsymbol{H}\boldsymbol{P}^f\boldsymbol{H}^T + \boldsymbol{R}\right]^{-1} \tag{2.10}$$

where $\boldsymbol{x}^a$ and $\boldsymbol{x}^f$ are the prior and posterior estimates of model states, respectively. $\boldsymbol{K}$ is the Kalman gain and $\boldsymbol{P}^f$ represents the background or forecast error covariance matrices of the model states. Since the true model states $\boldsymbol{x}_{\text{true}}$ are unknown, $\boldsymbol{P}^f$ is approximated by:

$$\boldsymbol{P}^f = \overline{\left(\boldsymbol{x}^f - \overline{\boldsymbol{x}}^f\right)\left(\boldsymbol{x}^f - \overline{\boldsymbol{x}}^f\right)^T} \tag{2.11}$$

where $\overline{\boldsymbol{x}}^f$ refers to the mean of the prior estimates $\boldsymbol{x}^f$.

## 2.3. Gamma Test theory

The Gamma test is normally used to estimate the variance of the noise in a given data set which is used to build the best smooth model, without knowledge of the specific model form. In this work, we will use the Gamma test to estimate the variance of the uncertainty between prior and posterior model states to correct posterior model states. In doing so, we assume that the uncertainty in particles is maintained. Only a brief introduction to the Gamma test theory is given in this subsection and further details can be found in corresponding papers [7, 8]. Let us assume we have prior particles $\boldsymbol{x}^{prior}$ and updated posterior particles $\boldsymbol{x}^{post}$:

$$\left\{\left(\boldsymbol{x}_i^{prior}, \boldsymbol{x}_i^{post}\right) \mid 1 \le i \le N_p\right\} \tag{2.12}$$

where $N_p$ is the number of particles.

In data assimilation, the prior particles are updated by observations and this process can be interpreted as a "data assimilation model" to generate the output-posterior particles according to the input-prior particles. Because of various uncertainty sources in data assimilation, such as the uncertainties caused by assumptions about observation error, forward operator error and observation bias [9], there always is an uncertainty in this "data assimilation model" that cannot be estimated. In this research, the variance of the uncertainty is estimated by the Gamma test. To fit the Gamma test, the relationship between $\boldsymbol{x}^{prior}$ and $\boldsymbol{x}^{post}$ is expressed as:

$$\boldsymbol{x}^{post} = F_{DA}\left(\boldsymbol{x}^{prior}\right) + \boldsymbol{r}_{DA} \tag{2.13}$$

where $F_{DA}$ represents a data assimilation process and $\boldsymbol{r}_{DA}$ are the errors with expectation zero. Particle filters are a non-Gaussian type of filter because when calculating weights of particles, the error distribution does not have to be Gaussian, but we still need to know the specific probability density function of the error distribution. However, in a nonlinear case, we would not be able to determine the error distribution. The Gamma test can estimate the variance of the noise var($\boldsymbol{r}_{DA}$) regardless of the specific data assimilation algorithm used and the underlying error distribution because of the existence of nonlinearity and non-Gaussianity.

The Gamma test statistic is calculated by the following procedure. First, the Gamma test uses a kd-tree to find the $k$th $(1 \le k \le p)$ nearest neighbors $\boldsymbol{x}_k^{prior}$ , $\boldsymbol{x}_k^{post}$ of $\boldsymbol{x}^{prior}$,$\boldsymbol{x}^{post}$ for each particle member. Here $p$ is set to 10 typically [8]. Next, the algorithm computes:

$$\delta\left(k\right) = \frac{1}{N_m}\sum_{i=1}^{N_m}\left|x_k^{prior} - x_i^{prior}\right|^2 \qquad (1 \le k \le p) \tag{2.14}$$

$$\gamma\left(k\right) = \frac{1}{2N_m}\sum_{i=1}^{N_m}\left|x_k^{post} - x_i^{post}\right|^2 \qquad (1 \le k \le p) \tag{2.15}$$

where |...| denotes Euclidean distance, $\delta\left(k\right)$ is the mean square of the $k$ nearest neighbors of the prior distribution, and $\gamma\left(k\right)$ is derived from the $k$ nearest neighbors of the posterior distribution, which is defined as the outcome of the Gamma Test. Based on the points $(\delta(k), \gamma(k))$, the linear regression $\gamma(k) = A\delta(k) + \Gamma$ is computed and the intercept $\Gamma$ is the estimation of $\mathrm{var}(r_{DA})$, as can be shown $\gamma(k) \to \mathrm{var}(r_{DA})$ in probability as $\delta(k) \to 0$.

## 2.4. Gaussian process regression

For a given model $\boldsymbol{f}(\boldsymbol{x})$, Gaussian process regression offers a solution to describe the relation between input and output. Like a Gaussian distribution, a Gaussian process can be defined by a mean function $m(\boldsymbol{x})$ and a covariance function $k(\boldsymbol{x}, \boldsymbol{x}')$. Thus, a Gaussian process is over functions, which is shown as follows:

$$\boldsymbol{f}(\boldsymbol{x}) \sim GP\left(m(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x}')\right) \tag{2.16}$$

in which $\boldsymbol{x}$ is the input of a model. For every input $x$, it has a corresponding value of function $\boldsymbol{f}(\boldsymbol{x})$, which represents an associated random variable.

For training a GP model, a specific mean function and a covariance function must be chosen, and the values of all hyperparameters in these two functions can be found by optimizing the log marginal likelihood of a GP model. A trained GP model can be used to make new predictions based on a test data set, which is not used for the training process by computing its posterior. Let $\boldsymbol{f}$ be the known values of functions(training data $\boldsymbol{x}$), and let us assume that $\boldsymbol{f}_*$ contains the unknown values(test data $\boldsymbol{X}$) for calculating the posterior. The joint distribution of $\boldsymbol{f}$ and $\boldsymbol{f}_*$ can be expressed as:

$$\begin{bmatrix} \boldsymbol{f}(\boldsymbol{x}) \\ \boldsymbol{f}_*(\boldsymbol{x}) \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{\mu}_* \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Sigma}_* \\ \boldsymbol{\Sigma}_*^{\mathsf{T}} & \boldsymbol{\Sigma}_{**} \end{bmatrix}\right) \tag{2.17}$$

$$\begin{aligned} \boldsymbol{\Sigma} &= k(\boldsymbol{x}, \boldsymbol{x}') \\ \boldsymbol{\Sigma}_* &= k(\boldsymbol{x}, \boldsymbol{X}) \\ \boldsymbol{\Sigma}_{**} &= k(\boldsymbol{X}, \boldsymbol{X}') \end{aligned} \tag{2.18}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\mu}_*$ are means of training data and test data respectively, and for covariance, $\boldsymbol{\Sigma}$ is the training data covariance and $\boldsymbol{\Sigma}_{**}$ is the test data covariance. $\boldsymbol{\Sigma}_*$ is used for the training and calculating test data covariance. The conditional distribution of $\boldsymbol{f}_*$ given $\boldsymbol{f}$ can be expressed as Eq.( 2.19).

$$\boldsymbol{f}_*|\boldsymbol{f} \sim N\left(\boldsymbol{\mu}_* + \boldsymbol{\Sigma}_*^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{f} - \boldsymbol{\mu}), \boldsymbol{\Sigma}_{**} - \boldsymbol{\Sigma}_*^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_*\right) \tag{2.19}$$

In data assimilation, after replacing observation operators with GPR models, transferring model states into observation space can be interpreted as using trained GPR models to predict a new data set. Therefore, the mentioned test $\boldsymbol{X}$ can be model states generated by propagating a model over time. In this context, the corresponding mean and variance of the estimation are expressed as follows:

$$\boldsymbol{\mu}_{\boldsymbol{X}} = m(\boldsymbol{X}) + k(\boldsymbol{x}, \boldsymbol{X})^{\mathsf{T}} k(\boldsymbol{x}, \boldsymbol{x}')^{-1} \left(m(\boldsymbol{x}) - m(\boldsymbol{X})\right) \tag{2.20}$$

$$\boldsymbol{\sigma}_{\boldsymbol{X}}^2 = k(\boldsymbol{X}, \boldsymbol{X}') - k(\boldsymbol{x}, \boldsymbol{X})^{\mathsf{T}} k(\boldsymbol{x}, \boldsymbol{x}')^{-1} k(\boldsymbol{x}, \boldsymbol{X}) \tag{2.21}$$

One advantage of the GPR is that it can estimate the model output and meanwhile give its uncertainty estimation, as shown in Eq.(2.21). In doing so, the observation error can be represented with a Gaussian distribution. When training GPR models, one observation must be transferred into a vector by adding an error. It is inevitable to introduce extra sampling errors. However, the uncertainty estimated by Eq.(2.21) represents the total error, including the observation error, sampling error and the GPR error. It should be noted that the replacement of observation operators can bring extra GPR errors into the Bayesian framework. It leads to the posterior distribution with a wider range, which can avoid over-confident results in Monte Carlo simulation [10]. The GPR models from the python package- scikit-learn are used in this work.

## References

[1] A. Doucet, N. de Freitas, and N. Gordon, *An Introduction to Sequential Monte Carlo Methods,* in *Sequential Monte Carlo Methods in Practice*, Statistics for Engineering and Information Science (Springer, New York, NY, 2001) pp. 3–14.

[2] J. D. Hol, T. B. Schon, and F. Gustafsson, *On Resampling Algorithms for Particle Filters,* in *2006 IEEE Nonlinear Statistical Signal Processing Workshop* (2006) pp. 79–82.

[3] J. S. Lui and R. Chen, *Sequential Monte Carlo methods for dynamic systems,* Journal of the American Statistical Association; Alexandria 93, 1032 (1998).

[4] H. Zhang, S. Qin, J. Ma, and H. You, *Using Residual Resampling and Sensitivity Analysis to Improve Particle Filter Data Assimilation Accuracy,* IEEE Geoscience and Remote Sensing Letters 10, 1404 (2013).

[5] S. Hong, M. Bolic, and P. M. Djuric, *An efficient fixed-point implementation of residual resampling scheme for high-speed particle filters,* IEEE Signal Processing Letters 11, 482 (2004).

[6] G. Evensen, *The Ensemble Kalman Filter: Theoretical formulation and practical implementation,* Ocean Dynamics 53, 343 (2003).

[7] D. Evans and A. J. Jones, *A proof of the Gamma test,* Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences 458, 2759 (2002).

[8] A. J. Jones, D. Evans, and S. E. Kemp, *A note on the Gamma test analysis of noisy input/output data and noisy time series,* Physica D: Nonlinear Phenomena 229, 1 (2007).

[9] P. L. Houtekamer and F. Zhang, *Review of the Ensemble Kalman Filter for Atmospheric Data Assimilation,* Monthly Weather Review 144, 4489 (2016).

[10] J. Zhang, W. Li, L. Zeng, and L. Wu, *An adaptive Gaussian process-based method for efficient Bayesian experimental design in groundwater contaminant source identification problems,* Water Resources Research 52, 5971 (2016).

# 3

# A Local Particle Filter Using Gamma Test Theory for High-Dimensional State Spaces

*Particle Filters are non-Gaussian filters, which means that the assumption that the error distribution of the ensemble should be Gaussian is unnecessary. Like EnKF, particle filters are based on the Monte Carlo approximation to represent the distribution of model states. It requires a substantial number of particles to approximate the probability density function of states in high-dimensional models, which is prohibitive for real applications. In order to overcome problems with high dimensionality, localization was applied in an Ensemble-type data assimilation system. This study combines the localization in LETKF with particle filters and proposes a new local particle filter with the model state space correction using Gamma Test theory for high-dimensional models. A series of tests with various parameter settings, including different the numbers of particles, observation intervals, localization scale, inflation factors, and observation operators, were used to evaluate the performance of this new method using a Lorenz model with 40 variables. Besides, the proposed filter was applied in the Lorenz model with 1000 variables to evaluate its performance in the model with higher dimensions. The results show that the local particle filter is stable and has considerable potential for complex higher dimensional models.*

This chapter is based on: Wang, Zhenwu, Rolf Hut, and Nick Van de Giesen. ”A Local Particle Filter Using Gamma Test Theory for High-Dimensional State Spaces.” *Journal of Advances in Modeling Earth Systems* 12.11 (2020): e2020MS002130.

## 3.1. Introduction

Numerical models are used to forecast and estimate model states in many fields including meteorology, ocean science and hydrology. Accounting for different sources of uncertainty to improve the accuracy of such models has drawn considerable attentions in recent decades [1]. Data assimilation(DA) provides a solution to combine information from model estimations and observations to achieve better prediction performance and quantify uncertainty [2–4]. DA has been applied widely in geoscience [5–12].

EnKF and its variants have been applied frequently [13–15]. EnKF relys on the assumption that the error distribution is Gaussian [16]. Particle filters(PFs) are another ensemble-type data assimilation algorithm, which is proposed by Gordon *et al.* [17] and has been used in many low-dimensional models [18]. Particle filters, just like the EnKF, use a Monte Carlo approximation in which a certain number of particles are used to represent the distribution of model states. The distribution reflects the mean and spread of model states and is updated by the prior weights which can be calculated using the likelihood given to each particle. However, unlike EnKF and its variants, PFs do not rely on the assumption that the error distribution is Gaussian. It should be noted that the Gaussian assumption can lead to suboptimal results when a model system is nonlinear and errors are non-Gaussian, which is the main limitation of ensemble-Gaussian-type data assimilation strategies.

For particle filters, the increasing number of dimensions of a model requires an exponentially growing number of particles to avoid filter divergence, as shown by Snyder *et al.* [19], which is called the curse of dimensionality [20]. Bengtsson *et al.* [20] also showed that for accurate representation of high dimensional distributions, the number of particles need to be increased exponentially with the number of model states. Therefore, the use of particle filters in high-dimensional models is limited due to the curse of dimensionality and associated high demand on computational resources.

Currently, there are several strategies to deal with the dimensionality of particle filters in high dimensional systems [21, 22]. An equivalent weights particle filter has been proposed, which prevents filter degeneracy using the proposal transition density [23, 24]. The proposal transition density keeps particles close to observations. Consequently, it leads to a better statistic representation of the posterior distribution because none of particles is ignored. Ades and van Leeuwen [25] explored the effect of the equivalent-weights particle filter on the dynamical balance in a primitive equation model. Their results showed that this method had potential for large-scale geophysical applications. Ades and van Leeuwen [26] used this methods in the barotropic vorticity model with 65500 states and the results illustrated its powerful abilities to make the filter stable and avoid filter divergence.

Localization is an useful method for preventing the collapse of PFs in high-dimension models, which is used commonly to solve the high-dimensional issues in EnKF and its variants, for example the LETKF [27]. Researchers have attempted to apply localization in particle filters to reduce the impact of dimensionality constraints. van Leeuwen [28] and Bengtsson *et al.* [29] tried to combine localization with particle filters. Farchi and Bocquet [30] reviewed several local particle filters from both theoretical and practical aspects. In this review, the localization solutions were divided into two strategies, which were state-domain localization and sequential-observation localization [30].

State-domain localization is used to update each model state independently only using the observations within localization scales, which has been applied in PFs [31–34]. Penny and Miyoshi [31] introduced a localized particle filter(LPF) with state-domain localization. The particles of this algorithm were updated first by a transformation matrix and the final updated particles were a linear combination of particles around each point. This LPF could outperform the LETKF under some conditions [31]. However, this method still needs an inflation method, which is a common method in data assimilation to maintain the spread of the model states to prevent collapse of the algorithm and to keep the filter stable.

In sequential-observation localization, all observations are assimilated one by one. Only nearby model states at each observation site are updated and influenced and distant model states stay unchanged [30]. Poterjoy [35] proposed a LPF with sequential-observation localization, which can be applied in high-dimensional systems using a small number of particles. It extended the weights in traditional particle filters to a vector to remove the impact of distant observations. Kernel density distribution mapping(KDDM) was applied to the updated particles to obtain a desired posterior distribution which was selected by prior particles and their weights. However, this method still suffered imbalance issues caused by localization. Poterjoy and Anderson [36] used an idealized atmospheric general circulation model (GCM) to test physical consistency of posterior particles and scalability. The results of these two different tests showed the LPF proposed by Poterjoy [35] is highly likely to work in more complex real models.

This study provides a new solution to prevent the divergence of particle filters in high-dimensional models. We attempt to consider the impact of the uncertainty between prior and posterior particles. Our new method is based on the Bayesian theorem and combines the standard particle filters with the so-called state-domain localization. Besides, the Gamma test is used to correct uncertainty in the posterior distribution. Localization divides all model states into local patch vectors and each model state has its own local patch vector [31]. In this way, particle filters only assimilate observations within the localization scale. The Gamma test is a technique which can be used to estimate uncertainty between prior and posterior model state space [37], which can lessen impact of localization issues. Additionally, the proposed algorithm applies the Gamma test to modify the variance of the posterior uncertainty to stabilize the filter and avoid filter collapse. The main motivation for designing this local filter is to make particle filters available and effective in high dimensional systems with a comparatively small number of particles.

The chapter is organized as follows. Section 3.2 describes the proposed algorithm in detail. Section 3.3 evaluates the performance of the proposed method using the Lorenz model with 40 and 1000 variables and compares its results with the LETKF. In the final section, the limitations of the new filter and its possible applications for high dimensional geophysical models are discussed.

## 3.2. Methodology

### 3.2.1. Degeneracy of filters

As introduced in section 2.1, particle filters do not rely on estimating the error distributions and useful particles are chosen based on weights using a resampling method.

The degeneracy of particle filters refers to the situation in which the particle with the largest weight is the only particle chosen by the resampling algorithm [35]. When an increase in the number of model states is not matched with an increase in the number of particles [19, 31], it becomes difficult for particle filters to find enough particles with sufficiently high probabilities. Increasing the number of particles can reduce the degeneracy of particle filters. Snyder *et al.* [19] analyzed the relationship between the number of particles and the number of model states theoretically. They indicated that the number of particles must increase exponentially with the dimensions of a model to obtain a posterior mean which has a smaller error than the prior particles or observations. LETKF has a different fundamental reason why the filter collapses. The main cause of the collapse of Particle filters is that the weights of each model state are close to unity. For LETKF, just like other EnKF variants, the classic divergence of a filter is indicated by the decreasing or increasing spread of ensemble [38]. The reason for this can be model errors, sample errors, nonlinearity in the system and other uncertainties.

### 3.2.2. Localization method

The localization method, which was proposed by Hunt *et al.* [27], is commonly used to remove spurious error covariance outside the local scale due to sampling errors caused by too small an ensemble. It only assimilates observations within a given scale for each particle, which is efficient for high-dimensional systems. The local particle filters in this study uses a localization scheme inspired by the LETKF. Therefore, the localization in LETKF is applied.

Using a localization method can stabilize the filter and avoid filter degeneracy in EnKF and its variants when only a small number of particles is used in the data assimilation of high-dimensional models [27]. For the issues in particle filters caused by dimensionality [19], as mentioned and explained in section 3.2.1, localization has been applied to solve the dimensionality of particle filters [31, 35]. However, the localization method can deal with the issue partially and additional methods are still needed to stabilize the posterior distribution. Poterjoy [35] corrected updated ensemble by using kernel density distribution mapping, and Penny and Miyoshi [31] still used an inflation method to prevent the filter collapse. As in LETKF, every model state is assimilated one by one in particle filters and particles and observations are localized by a localization function to form local particles $\boldsymbol{x}_{loc}$ and local observations $\boldsymbol{y}_{loc}$. Then the weights of each state are calculated based on $\boldsymbol{y}_{loc} - \mathrm{H}\boldsymbol{x}_{loc}$. In this study, the localization method in LETKF is used for particle filters to remove certain observations that fall outside the localization scale. The specific localization function used here for local particle filters was proposed by Gaspari and Cohn [39].

### 3.2.3. Local particle filters with the Gamma test

Chapter 2 gives the introduction to the standard particle filters in Section 2.1 and Gamma test theory in Section 2.3. The local particle filter with the Gamma test(LPF-GT) is explained in this subsection. The foundation of particle filters is still Bayes' theorem and the Monte Carlo method. With the localization procedure, each model state can be updated independently in DA. The Gamma test provides an estimation of potential uncertainty for DA. Under the assumption that the observation errors are independent, weights can be calculated by:

$$\boldsymbol{w}_n = p\left(\boldsymbol{y} \mid \boldsymbol{x}_n\right) = \prod_{i=1}^{N_{obs}} p\left(y_i \mid \boldsymbol{x}_n\right) \tag{3.1}$$

where $N_{\text{obs}}$ is the number of observations. When calculating weights in PFs, and when only a few particles carry almost all the weight, filter collapse is inevitable. Therefore, in order to avoid filter collapse, the probability $p\left(y_i \mid \boldsymbol{x}_n\right)$ in Eq.(3.1) is calculated in the following way:

$$p_\beta\left(y_i \mid \boldsymbol{x}_n\right) = \left[p\left(y_i \mid \boldsymbol{x}_n\right) * loc\left(y_i, \boldsymbol{x}_n, r\right) + \beta_a\right] \beta_m \tag{3.2}$$

where $loc\left(y_i, \boldsymbol{x}_n, r\right)$ are the localization coefficients and $*$ represents elementwise product. In current research, we use (4.10) of Gaspari and Cohn [39] for $loc\left(y_i, \boldsymbol{x}_n, r\right)$, which has a Gaussian-type structure with a width $r$. The parameters $\beta_a$ and $\beta_m$ are used to control weights. Hence, after replacing $p\left(y_i \mid \boldsymbol{x}_n\right)$ with $p_\beta\left(y_i \mid \boldsymbol{x}_n\right)$, the localized weights in LPF-GT are given by:

$$\begin{aligned} \boldsymbol{w}_n &= \prod_{i=1}^{N_{obs}^{loc}} p_\beta\left(y_i \mid \boldsymbol{x}_n\right) \\ &= \prod_{i=1}^{N_{obs}^{loc}} \left[p\left(y_i \mid \boldsymbol{x}_n\right) * loc\left(y_i, \boldsymbol{x}_n, r\right) + \beta_a\right] \beta_m \end{aligned} \tag{3.3}$$

where $N_{obs}^{loc}$ indicates the number of observations within the localization scale. $\beta_a$ is an additive factor and $\beta_m$ is a multiplicative factor.

The mean effective number of particles $N_{eff}$(shown in Eq.(3.4)) is used to evaluate the quantity of the particles [31] and the factor $\beta_a$ and $\beta_m$ in Eq.(3.3) can tune the value of $N_{eff}$.

$$N_{eff} = \left[\sum_{i=1}^{N_p} \left(\frac{w_i}{\sum_{i=1}^{N_p} w_i}\right)^2\right]^{-1} \tag{3.4}$$

The role of $\beta_a$ and $\beta_m$ is mainly to avoid filter degeneracy. Because they influence the value of $N_{eff}$. Tuning $N_{eff}$ can change the percentage of removed particles in the resampling procedure, which was discussed in the next section. We found that keeping $N_{eff}$ close to a certain value by tuning $\beta_a$ and $\beta_m$ can avoid filter collapse. To fix $N_{eff}$, values of $\beta_a$ and $\beta_m$ are changed dynamically by finding appropriate values in the parameter space of $\beta_a$ and $\beta_m$. There are definitely more than more than one pair of $\beta_a$ and $\beta_m$ that fix $N_{eff}$. We simply stopped the search when the first pair of these two parameters had been found.

Eq.(3.3) attempts to avoid the divergence of particle filters by rescaling the weights of each particle twice. The weights of distant observations are reduced gradually by introducing localization coefficients to the probabilities of local observation errors. The factors $\beta_a$ and $\beta_m$ are used to maintain the stability of the weights, which can control the proportion of particles removed by residual resampling algorithm.

Next, the modified particles $\boldsymbol{x}_n^a$ and the factor $\alpha$ corrected by the Gamma test are given in Eq.(3.5) and Eq.(3.9) respectively.

$$\boldsymbol{x}_n^a = \overline{\boldsymbol{x}_n^{a'}} + \eta\left(\boldsymbol{x}_n^{a'} - \overline{\boldsymbol{x}_n^{a'}}\right) + (1-\eta)\,\alpha\left(\boldsymbol{x}_n^{a'} - \boldsymbol{x}_n^b\right) \tag{3.5}$$

where $\boldsymbol{x}_n^{a'}$ indicates the updated particles generated by a resampling method. The prime indicates a posterior, or updated, particle. And $\overline{\boldsymbol{x}_n^{a'}}$ is its mean. $\boldsymbol{x}_n^b$ denotes the prior particles. In Eq.(3.5), $\alpha\left(\boldsymbol{x}_n^{a'} - \boldsymbol{x}_n^b\right)$ represents the uncertainty $\boldsymbol{r}_{DA}$ brought by data assimilation, which is shown in Eq.(2.13). We assume that $\boldsymbol{r}_{DA}$ follows a Gaussian distribution with mean zero and its variance is estimated by the Gamma test. Using $\boldsymbol{x}_n^{a'}$ to minus $\boldsymbol{x}_n^b$ directly, we obtain a sample, which is not the error we expect. To achieve our goal, the sample $\left(\boldsymbol{x}_n^{a'} - \boldsymbol{x}_n^b\right)$ needs to be normalized, which is given as $\boldsymbol{X}$. Next, We set the variance of $\boldsymbol{X}$ equal to $\Gamma$ estimated by the Gamma test. According to the definition of the variance, we define $\Gamma$ and normalized $\mathrm{var}\left(\boldsymbol{x}_n^{a'} - \boldsymbol{x}_n^b\right)$ as follows.

$$\Gamma = \frac{\sum(\alpha\boldsymbol{X} - \mu)^2}{N} \tag{3.6}$$

$$\mathrm{var}\left(\boldsymbol{x}_n^{a'} - \boldsymbol{x}_n^b\right) = \frac{\sum(\boldsymbol{X} - \mu)^2}{N} \tag{3.7}$$

where $\Gamma$ is the desired variance and $\alpha$ refers to the corrected factor of the normalized sample $\boldsymbol{X}$. The mean of $\boldsymbol{X}$ is zero. $N$ is the number of samples. We can obtain $\alpha$ in Eq.(3.5) based on Eq.(3.8) and Eq.(3.9).

$$\frac{\Gamma}{\mathrm{var}\left(\boldsymbol{x}_n^{a'} - \boldsymbol{x}_n^b\right)} = \frac{\sum(\alpha\boldsymbol{X} - \mu)^2}{\sum(\boldsymbol{X} - \mu)^2} = \frac{\alpha^2\sum(\boldsymbol{X})^2}{\sum(\boldsymbol{X})^2} = \alpha^2 \tag{3.8}$$

$$\alpha = \sqrt{\frac{\Gamma}{\mathrm{var}\left(\boldsymbol{x}_n^{a'} - \boldsymbol{x}_n^b\right)}} \tag{3.9}$$

In Eq.(3.5), the uncertainty of posterior particles is a linear combination of the particles updated by particle filters $\left(\boldsymbol{x}_n^{a'} - \overline{\boldsymbol{x}_n^{a'}}\right)$ and particles modified by the Gamma test $\alpha\left(\boldsymbol{x}_n^{a'} - \boldsymbol{x}_n^b\right)$. The mean of $\boldsymbol{x}_n^a$ is still obtained using the resampling algorithm. In this research, we assume that the uncertainty of $\boldsymbol{x}_n^a$ consists of two parts. One part is comes from the resampling method. The other part comes from the data assimilation framework, which can be estimated by the Gamma test. These two parts are combined by $\eta$ and the value of $\eta$ is between 0 and 1. The parameter $\eta$ can be tuned to change the impact on the uncertainty of $\boldsymbol{x}_n^a$.

**Algorithm 1** Pseudocode description of LPF-GT

```
for $n = 1 \rightarrow N_m$ do
    call Localization to obtain local index $id_{\text{loc}}$ and local coefficients $c_{\text{loc}}$ and then get the local observations $(y_{i,\text{local}})$ and local particles $x_{\text{local}}^{(y_{i-1})}$
    for $j = 1 \rightarrow N_p$ do
        $w_{n,j} \leftarrow p\left(x_{j,\text{local}}^{(y_{i-1})} \mid y_{i,\text{local}}\right)$
    end for
    call a bisection function to find factors $\beta_a$ and $\beta_m$, which brings $N_{eff}$ close to a certain value
    $w_n \leftarrow \prod \left(w_{n,j} + \beta_a\right) * \beta_m$
    $w_{n,\text{nor}} = \frac{w_n}{\sum w_n}$
    Obtain resampled particles $x_n^{(y_i)}$ based on $w_{n,\text{nor}}$
end for
call GammaTest($x^{(y_i)}, x^{(y_{i-1})}$) to obtain $\Gamma$
$\alpha = \sqrt{\frac{\Gamma}{\text{variance}\left(\text{x}^{(\text{y}_\text{i})}, \text{x}^{(\text{y}_{\text{i}-1})}\right)}}$
$\bar{x} = \text{mean}\left(x^{(y_i)}\right)$
$x^{(y_i)} = \bar{x} + \eta\left(\bar{x} - x^{(y_i)}\right) + (1-\eta)\,\alpha\left(x^{(y_i)} - x^{(y_{i-1})}\right)$, where $\eta \in [0, 1]$

function GammaTest($x^{(y_i)}, x^{(y_{i-1})}$)
    calculate the Gamma test statistic $\Gamma$
    return $\Gamma$
end function

function Localization
    calculate local index $id_{\text{loc}}$ and local coefficients $c_{\text{loc}}$
    return $id_{\text{loc}}, c_{\text{loc}}$
end function
```

## 3.3. Numerical experiments and results

In this research, the Lorenz model [40, L96 model] was used to evaluate our proposed local particle filters(LPF-GT) and to compare its performance to LETKF. More detailed introduction to the L96 model can be found in Lorenz [40]. As is common in testing new data assimilation schemes for high dimensions models, in this study, 40 variables are chosen and $F$ remains fixed at 8.0 to maintain the chaotic behavior in L96. Furthermore, the differential equations of the L96 model are integrated by a fourth-order Runge–Kutta method with a 0.05 time step [40, which is defined as 6h].

### 3.3.1. Experimental setup

A set of experiments were conducted to test the validity of LPF-GT for various parameter configurations that mimic real applications. In these experiments, LPF-GT and LETKF were examined by a variety of parameter settings to test the effectiveness and disadvantages of LPF-GT. The parameters in these experiments include various observation intervals, the number of particles $N_p$, the number of observations $N_{obs}$, inflation factors, localization scales, and two different observation operators $H$.

The default configuration consists of the linear $H$ operator. Only half of the model states are observed, 20 observations, were assimilated in every experiment which were chosen evenly and were fixed spatially over time. A inflation methods was used for LETKF,and the inflation factor had been tuned to a fixed value 1.05 as the default. The parameters $\beta_a$ and $\beta_m$ in LPF-GT have a similar role as in the inflation method. The influence to $N_{eff}$ by tuning $\beta_a$ and $\beta_m$ was investigated through numerical experiments. The default number of particles was 100. Observations were derived from the truth with an uncertainty $\epsilon \sim N\,(0, 0.5)$. Other configurations, which are different from the default, will be given at the beginning of each experiment. A spinup time with 1000 time steps was added to each test and the following 10000 steps were used to summarize and analyze results. All experiments were executed on the DAS-4 supercomputer [41].

In all experiments, the root mean square error(RMSE) was used to evaluate the performance of the new filter. The ensemble spread, defined as the square-root of the variance of the ensemble averaged over all model states, is another metric for the evaluation. These two metrics are defined as follows.

$$\mathrm{RMSE} = \sqrt{\frac{1}{N_m}\sum_{k=1}^{N_m}\left(x^t_{\mathrm{truth},k} - \bar{x}^t_k\right)^2} \tag{3.10}$$

$$\mathrm{Spread} = \sqrt{\frac{1}{N_m}\sum_{k=1}^{N_m}\sum_{i=1}^{N_p}\left(x^t_{i,k} - \bar{x}^t_k\right)^2} \tag{3.11}$$

in which $\bar{x}^t_k$ is the ensemble mean for filters and $x^t_{\mathrm{truth},k}$ is the corresponding truth. $x^t_{i,k}$ represents each particle. In this research, we use time-averaged values for both metrics.

### 3.3.2. Results

The first experiment was to examine the behavior of one model state, and the performance of the local filter when a linear $H$ operator was used in the simulation. The results of prior error statistics from LETKF and LPF-GT were compared by calculating domain averages of RMSE and ensemble spreads for all model states over time. Time series of the first model state and the corresponding truth have been plotted together in Fig.3.1. The constants $\eta$ and $N_{eff}$ in LPF-GT can be tuned, and using different combinations can impact the final performance. In this experiment, we used $\eta = 0.55$ and $N_{eff} = 0.65$, and the behavior of the system for these two parameters was investigated later. For LETKF, after tuning the inflation factor, the fixed value for inflation 1.05 was used. In this linear-Gaussian case, at the beginning of data assimilation, LPF-GT took more time than LETKF to reach a stable status. It is probably because LPF-GT is more sensitive to sampling errors. When both of them become stable, they produce low RMSEs and have similar performance. For the entire simulation time, results show that LETKF outperforms LPF-GT. The time-averaged RMSE and ensemble spread for LPF-GT are 0.38 and 0.55, respectively, compared to 0.17 and 0.29 for LETKF. This experiment demonstrates that LPF-GT can work stably for high-dimensional models using the linear observation operator and confirms that the application of localization in particle filters prevents the filter collapse using fewer particles.

Next, we explore the impact of $\eta$ and $N_{eff}$, when different localization scales $v_{local}$ are used. Therefore, a set of experiments using the linear $H$ operator for different combinations of these parameters were conducted, and the RMSE and spread of prior particles averaged over the entire domain are shown in Fig.3.2. LPF-GT is tuned optimally in this case, and the optimal configuration comes from the experiment which yields the lowest prior RMSEs. Four localization scales $v_{local}$ used in this test are 1.0, 5.0, 10.0 and 15.0 respectively. The tested values for the parameter $\eta$ were 0.45, 0.5 and 0.55 and parameter $N_{eff}$ varies between 0.4 and 0.65 with 0.05 steps. Each pixel is the result of running the data assimilation algorithm with different settings of these three parameters. For the current settings $\eta = 0.5$, $N_{eff} = 0.65$ and $v_{local} = 5$ yielded the best result.

These results in Fig.3.2 clearly show that changes in these two parameters, $N_{eff}$, and $\eta$ can impact the performance, and both of their roles are significant. When $v_{local} = 5$ is used, the RMSE becomes lower with the growth of $N_{eff}$. In cases with larger localization scales, in general, the performance of the LPF-GT becomes worse with increasing $N_{eff}$. As for the impact of the localization, increasing localization scales degrade the accuracy of the LPF-GT generally, which is consistent with the results in Penny and Miyoshi [31]'s research. The $\eta$ parameter is used to adjust the analysis errors, which are derived from analysis errors given by the PF and the other one corrected by the Gamma test, but there is no apparent monotonic relationship between the two error sources. Therefore, appropriate parameter values of $\eta$ should be tuned to obtain the desired performance. The parameter $N_{eff}$ influences the performance of LPF-GT by reassigning values of weights. Using a proper value of $N_{eff}$ can draw model states to the truth, and it turns out that changing the variance of particles by adjusting their weights is a potential strategy to avoid the collapse for particle filters.

Next, we check the sensitivity of the proposed LPF-GT to the number of particles and the impact of the observation assimilation intervals in both linear and nonlinear

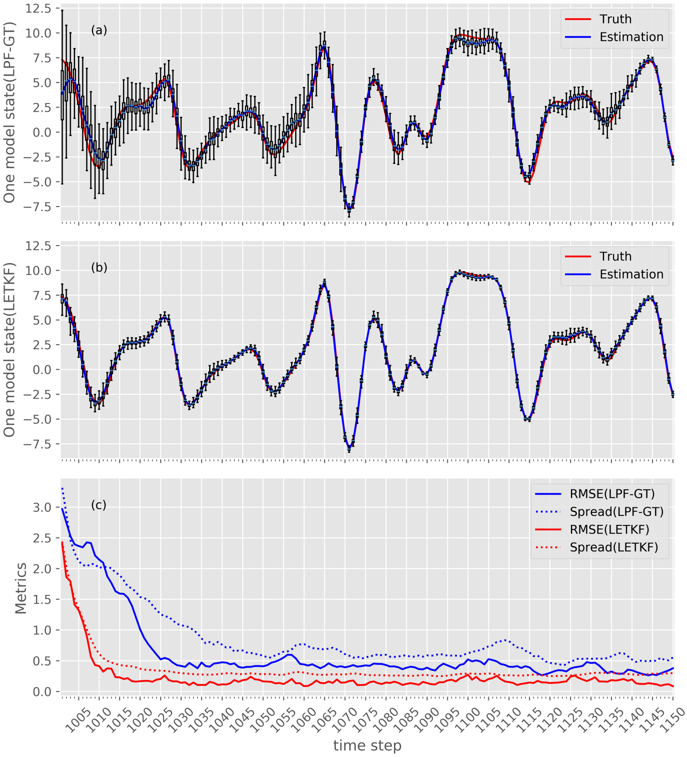

Figure 3.1: Trajectories of (a) the first model state using the LPF-GT, (b) the first model state using the LETKF and (c) RMSE and spread for both methods. In the first two figures, the blue curve is the truth and the red curve is the estimation of the model state. In the third one, the solid lines indicate the RMSE and the dash lines is the spread. The assimilation starts at time step 1,000.

cases. In the nonlinear case, $y = \ln|x|$ is used as the nonlinear observation operator. In this research, the number of particles varies as 50, 100, 200, 300 and 500. The standard deviation of the error in observation is set to 0.5 in the linear case, and both 0.1 and 0.5 are chosen as observational errors in the nonlinear case.

Fig.3.3 shows the RMSE, and the corresponding ensemble spread as a function of the number of particles and update interval in the linear case. For the linear/Gaussian experiment, both filters produce low RMSEs and the LETKF has lower RMSEs than the LPF-GT but from resulting prior statistics the performance of the LPF-GT is still acceptable. LPF-GT is more sensitive to the number of particles and, as expected, the RMSEs decrease when the number of particles is increased. For both filters, using 50 particles or ensemble members can provide satisfactory results. The LETKF is optimal for the linear observation type, and because of the application of localization, it does not need a large number of ensemble members to maintain the Gaussianity of the ensemble. Performance is limited by sampling errors in prior ensembles.

In Fig.3.3, the spread given by LPF-GT is larger than results from LETKF. It is mainly because we use several methods to avoid the collapse of the filter including

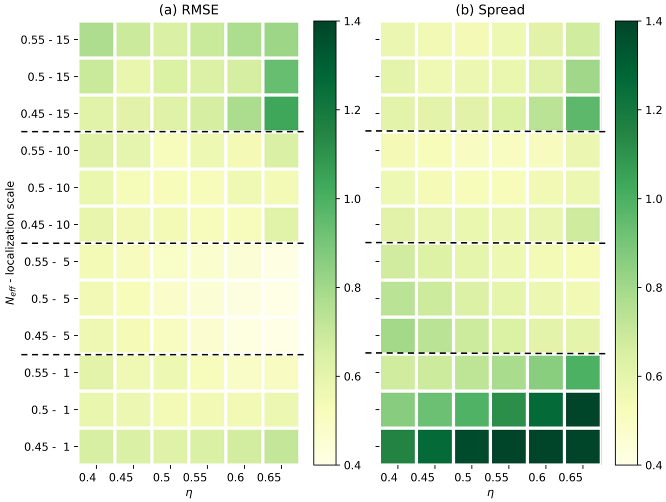

Figure 3.2: Prior mean RMSE (a) and spread (b) as a function of parameter $N_{eff}$, $\eta$ and localization scale $v_{local}$ . The value in brackets indicates the localization scale and each cell represents one experiment.

the localization procedure, tuning the effective number of particles and considering potential uncertainty provided by the Gamma test. All these methods keep the filter stable, but the variance of particles is inflated inevitably.

To examine the performance of the LPF-GT when nonlinear observations are assimilated, a more comprehensive assessment of its performance is given. The experimental configurations for the nonlinear operator include two localization scales and two observation errors. Fig.3.4 shows the prior mean RMSE and the corresponding spread as a function of the number of particles in the nonlinear case, when using various localization radius and observational errors. The nonlinear $H$ operator was used for the nonlinear experiments. However, because the LETKF is suboptimal for the nonlinear operator: $y = \ln|\boldsymbol{x}|$, its performance is not as stable and predictable as LPF-GT and the filter degeneracy occurs in LETKF in many experiments. Thus, in Fig.3.4, we only demonstrate results from the LETKF when the collapse did not happen.

From results in Fig.3.4, it is clear that LPF-GT can provide more accurate solutions than the LETKF under these four conditions and the non-Gaussianity introduced by the nonlinear measurement operators makes LETKF less effective. When observational errors with standard deviation 0.1 are used, LPF-GT only needs 50 particles to work stably and continuously, and LETKF only outperforms LPF-GT slightly with 200 particles. In the case with standard deviation 0.5, LPF-GT requires 100 particles to achieve a relatively acceptable performance. With limited ensemble members, LETKF needs to concentrate on the first two moments of the posterior distribution and these two moments will not be unbiased in the nonlinear case. The additional nonlinearity makes LETKF more sensitive to changes of parameters like ensemble sizes and localization scales, which accounts for the collapse of LETKF in these cases partially. The LPF-GT can maintain its ability and stability with relatively high accuracy even with different

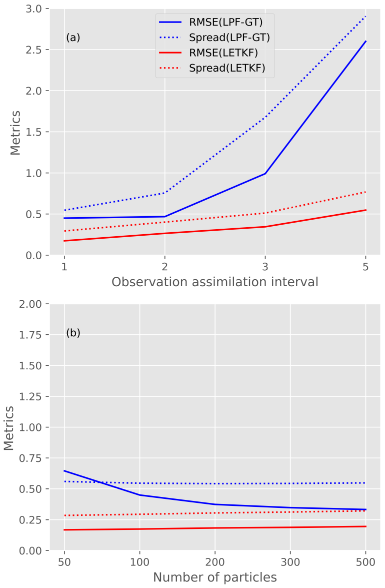

Figure 3.3: Prior mean RMSE and spread as a function of (a) ensemble size and (b) assimilation time interval for experiments using the linear observation operator.

localization radius, and observational errors, which produces significant improvements over LETKF. Meanwhile, these experiments also provide useful information about the impact of localization scales. For LPF-GT, its performance becomes better when the localization scale is increased to five.

The performance of LPF-GT is also investigated for cases in which observations are assimilated at different frequencies. The timescale $dt = 0.05$ of the Lorenz model was applied in this research, and it is comparable to the error doubling happening over six hours in the operational forecasting systems [40]. Values of 1, 2, 3, and 5 are used for different time intervals, which represent 6h, 12h, 36h, and 72h update frequencies, respectively.

Similar to experiment settings for the number of particles, we ran simulations in both linear and nonlinear cases using different time intervals. The results for the linear operator are shown in Fig.3.5. From simulation results, the LPF-GT has no practical benefit over LETKF when observations are assimilated less frequently. Using a longer update time interval can accumulate model errors, which is the main reason why the performance of both filters becomes worse when increasing the time interval. When using the nonlinear operator, LPF-GT offers advantages over the LETKF, which is shown in Fig.3.5.

When using localization radius five and observational errors with standard deviation 0.5, LETKF achieves a somewhat better performance than LPF-GT. However, in other cases, LPF-GT offers substantial benefits over LETKF, which shows that LETKF is more easily influenced by the localization radius. Probably, searching more of the parameter space for LETKF can produce better results, but it increases the cost. Besides, similar

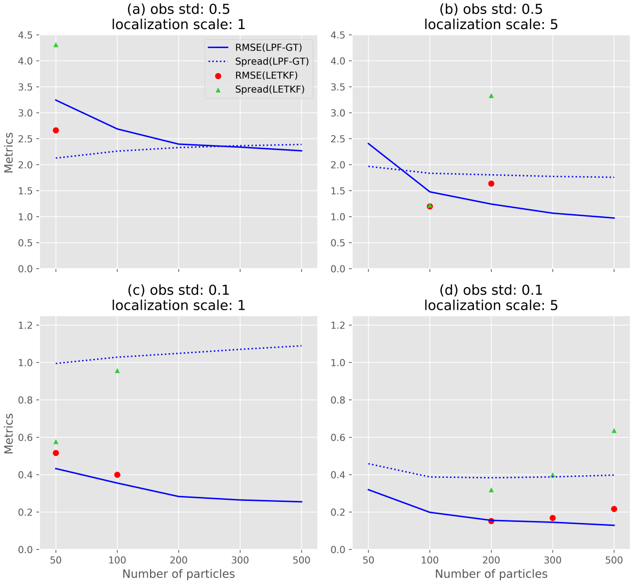

Figure 3.4: Prior mean RMSE and spread as a function of ensemble size for (a) obs std (the standard deviation of observational error) 0.5 and localization scale 1, (b) obs std 0.5 and localization scale 5, (c) obs std 0.1 and localization scale 1, (d) obs std 0.1 and localization scale 5, using the nonlinear observation operator $y = \ln|x|$.

to results in experiments of the number of particles, LPF-GT yields better results when the localization scale five used.

Since the performance of LPF-GT in high-dimensional models is of high interest, we investigate the behavior of LPF-GT in a non-linear case using the Lorenz model with 1000 variables. We used fewer particles to explore the performance limit of LPF-GT. Half of all model states were observed, and non-linear observations generated from the truth and the operator $y = \ln|x|$ with the observation error $\epsilon \sim N(0, 0.1)$ were assimilated by using 25, 35, and 50 particles in the experiments. The rest of the experimental settings remained the same.

Fig.3.6 shows average prior mean RMSEs and spread for the different numbers of particles. For the case without data assimilation in this research, when a number of particles is propagated in the Lorenz model over time, the average mean RMSE of particles is about 3.5. It means that when the value of RMSE is smaller than 3.5, data assimilation improves model estimations. Otherwise, data assimilation is likely to reduce the accuracy of the model. From the results in Fig.3.6, it becomes clear that applying LETKF to the non-linear case, causes filter collapse, and its domain-averaged prior RMSEs in all experiments are close to 5.0. For LPF-GT, increasing the number of particles improves the performance of data assimilation as expected. In the case with 25 or 35 particles, we can conclude that LPF-GT helps improve the accuracy of the model. But to achieve a satisfactory result, LPF-GT needs at least 50 particles.

To investigate the stability of filters, time series of the second model state, which is unobserved, and the development of the spread and the RMSE of particles in the

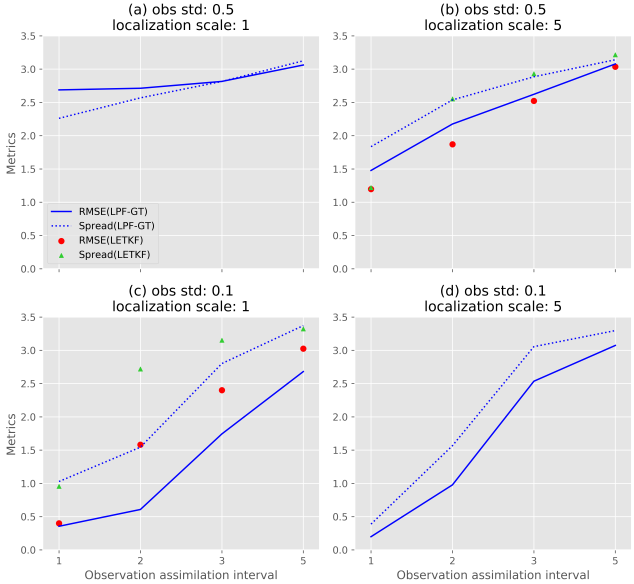

Figure 3.5: Prior mean RMSE and spread as a function of assimilation time interval for (a) obs std (the standard deviation of observational error) 0.5 and localization scale 1, (b) obs std 0.5 and localization scale 5, (c) obs std 0.1 and localization scale 1, (d) obs std 0.1 and localization scale 5, using the nonlinear observation operator $y = \ln|x|$.

experiment using 50 particles, are shown in Figure 3.7. The time series starts from the beginning of data assimilation, and it needs some time to stabilize. The first two rows show the comparison of the truth and the estimation of the second model state for LPF-GT and LETKF. LPF-GT gives a more accurate result and shows more stability than LETKF. The estimation of LETKF follows the truth in a short time, but after that, it deviates from the truth. The performance of LPF-GT is much more stable, and its estimations always stay close to the truth. From the development of the RMSE and the spread over time, it is clear that LETKF collapses at the beginning. By contrast, the spread and the RMSE of LPF-GT gradually decrease until the filter becomes stable. In this case, LPF-GT completely outperforms LETKF.

## 3.4. Conclusions

This research proposes a local PF for nonlinear high-dimensional applications. Similar to the localization method used in LETKF, LPF-GT assimilates observations within the localization scale, and the influence of distant observations are decreasing gradually in large-scale geophysical systems. Because of the use of the localization method, in this method, each model state needs much fewer observations than what is needed by typical particle filters. The LPF-GT updates particles sequentially when observations are available. Posterior weights for each particle are obtained based on the localized likelihood of observations for each state in a model. Particles with lower weights are removed by the resampling method. The mean of new particles is based on the re-

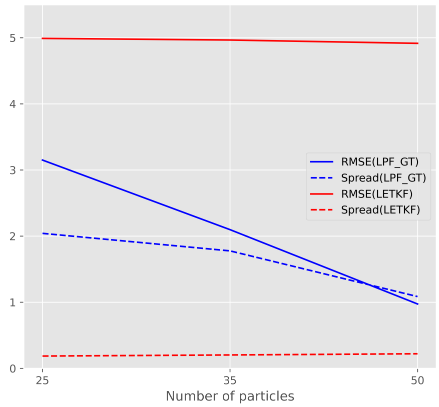

Figure 3.6: Prior mean RMSE and spread as a function of ensemble size for the standard deviation of observational error 0.1 and localization scale 5 using the nonlinear observation operator $y = \ln|\mathbf{x}|$ and 1000 particles.

sampled particles, and its uncertainty is a linear combination of sampled particles and the uncertainty estimated by the Gamma test. The proposed filter can prevent filter collapse, even when the number of particles is relatively small. Another advantage of the use of the localization method is to make computation of the new filter affordable for large applications. For each state, all calculations made are within the localization scales and can be parallelized easily. Correction by the Gamma test needs more computing time. However, for even larger systems, the computational cost of the Gamma test does not increase because the estimate of variance approaches a constant with an increase in the number of samples. It is unnecessary to use all samples for the Gamma test, and only a small number of samples is enough for the estimation of the variance.

The LPF-GT algorithm proposed in this research can be problematic for some applications in geoscience because of the imbalance caused by localization in model states, which is a disadvantage of the localization method. Similar to other data assimilation methods with localization, like LETKF and localized EnKF, LPF-GT can also break the physical consistency and cause an imbalance in posterior states. Poterjoy [35] found the imbalance issue when updated using their local PFs. The local PFs proposed by Penny and Miyoshi [31] showed a certain level of imbalance when they decreased the localization scale. The imbalance issue is common in most data assimilation methods with localization, and a proper localization scale is always needed for a specific application. One possible solution is to tune localization scales, but the cost of it may be expensive for some larger models. Thus, developing an adaptive localization method deserves more attention.

The LPF-GT has been tested by using the Lorenz system with 40 and 1000 variables. All results from a set of experiments show that the new filter is stable and avoids filter degeneracy successfully. In the ideal case with the linear observation operator, LETKF outperforms LPF-GT slightly. For nonlinear cases, LPF-GT provides a significant benefit over LETKF. The successful application of LPF-GT with a Lorenz model does not

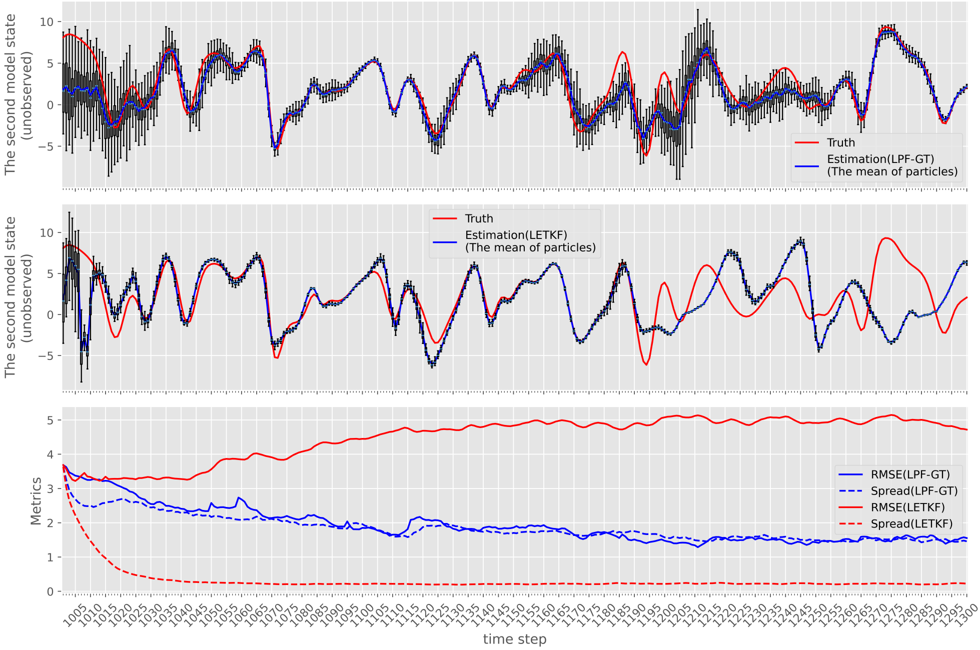

Figure 3.7: Time series of (a) the second model state using LPF-GT, (b) the second model state using LETKF and (c) RMSE and spread for both methods. In this experiment, the standard deviation of observational error and localization scale are set to 0.1 and 5 respectively and the nonlinear observation operator $y = \ln|\mathbf{x}|$ and 1000 particles are used. In the first two figures, the blue curve is the truth and the red curve is the estimation of the model state. In the third one, the solid lines indicate the RMSE and the dash lines is the spread. The assimilation starts at time step 1,000.

guarantee its success in other models. Therefore, its application in real world models with high dimensions will be the main topic of future studies to explore the limitations and advantages of this new filter.

## References

[1] Y. Liu, A. H. Weerts, M. Clark, H.-J. Hendricks Franssen, S. Kumar, H. Moradkhani, D.-J. Seo, D. Schwanenberg, P. Smith, A. I. J. M. van Dijk, N. van Velzen, M. He, H. Lee, S. J. Noh, O. Rakovec, and P. Restrepo, *Advancing data assimilation in operational hydrologic forecasting: Progresses, challenges, and emerging opportunities,* Hydrol. Earth Syst. Sci. 16, 3863 (2012).

[2] R. N. Bannister, *A review of operational methods of variational and ensemble-variational data assimilation,* Quarterly Journal of the Royal Meteorological Society 143, 607 (2017).

[3] P. J. van Leeuwen, *Nonlinear Data Assimilation for high-dimensional systems,* in *Nonlinear Data Assimilation*, Frontiers in Applied Dynamical Systems: Reviews

and Tutorials, edited by P. J. Van Leeuwen, Y. Cheng, and S. Reich (Springer, 2015) pp. 1–73.

[4] R. Hut, B. A. Amisigo, S. Steele-Dunne, and N. van de Giesen, *Reduction of Used Memory Ensemble Kalman Filtering (RumEnKF): A data assimilation scheme for memory intensive, high performance computing,* Advances in Water Resources Data Assimilation for Improved Predictions of Integrated Terrestrial Systems, 86, Part B, 273 (2015).

[5] Y. Choi, D.-H. Cha, M.-I. Lee, J. Kim, C.-S. Jin, S.-H. Park, and M.-S. Joh, *Satellite radiance data assimilation for binary tropical cyclone cases over the western North Pacific,* Journal of Advances in Modeling Earth Systems 9, 832 (2017).

[6] M. Fang and X. Li, *An Artificial Neural Networks-Based Tree Ring Width Proxy System Model for Paleoclimate Data Assimilation,* Journal of Advances in Modeling Earth Systems 11, 892 (2019).

[7] A. M. Fox, T. J. Hoar, J. L. Anderson, A. F. Arellano, W. K. Smith, M. E. Litvak, N. MacBean, D. S. Schimel, and D. J. P. Moore, *Evaluation of a Data Assimilation System for Land Surface Models Using CLM4.5,* Journal of Advances in Modeling Earth Systems 10, 2471 (2018).

[8] C. Irrgang, J. Saynisch, and M. Thomas, *Utilizing oceanic electromagnetic induction to constrain an ocean general circulation model: A data assimilation twin experiment,* Journal of Advances in Modeling Earth Systems 9, 1703 (2017).

[9] J. Dong, S. C. Steele-Dunne, J. Judge, and N. van de Giesen, *A particle batch smoother for soil moisture estimation using soil temperature observations,* Advances in Water Resources 83, 111 (2015).

[10] J. Jin, H. X. Lin, A. Segers, Y. Xie, and A. Heemink, *Machine learning for observation bias correction with application to dust storm data assimilation,* Atmospheric Chemistry and Physics 19, 10009 (2019).

[11] Y. Zhang, J. Hou, J. Gu, C. Huang, and X. Li, *SWAT-Based Hydrological Data Assimilation System (SWAT-HDAS): Description and Case Application to River Basin-Scale Hydrological Predictions,* Journal of Advances in Modeling Earth Systems 9, 2863 (2017).

[12] J. Jin, H. X. Lin, A. Heemink, and A. Segers, *Spatially varying parameter estimation for dust emissions using reduced-tangent-linearization 4DVar,* Atmospheric Environment 187, 358 (2018).

[13] L. Sun, O. Seidou, I. Nistor, and K. Liu, *Review of the Kalman-type hydrological data assimilation,* Hydrological Sciences Journal 61, 2348 (2016).

[14] X. Xie and D. Zhang, *Data assimilation for distributed hydrological catchment modeling via ensemble Kalman filter,* Advances in Water Resources 33, 678 (2010).

[15] H. Chen, D. Yang, Y. Hong, J. J. Gourley, and Y. Zhang, *Hydrological data assimilation with the Ensemble Square-Root-Filter: Use of streamflow observations to update model states for real-time flash flood forecasting,* Advances in Water Resources 59, 209 (2013).

[16] G. Evensen, *The Ensemble Kalman Filter: Theoretical formulation and practical implementation,* Ocean Dynamics 53, 343 (2003).

[17] N. J. Gordon, D. J. Salmond, and A. F. M. Smith, *Novel approach to nonlinear/non-Gaussian Bayesian state estimation,* IEE Proceedings F (Radar and Signal Processing) 140, 107 (1993).

[18] P. J. van Leeuwen, *Particle Filtering in Geophysical Systems,* Monthly Weather Review 137, 4089 (2009).

[19] C. Snyder, T. Bengtsson, P. Bickel, and J. Anderson, *Obstacles to High-Dimensional Particle Filtering,* Monthly Weather Review 136, 4629 (2008).

[20] T. Bengtsson, P. Bickel, and B. Li, *Curse-of-Dimensionality Revisited: Collapse of the Particle Filter in Very Large Scale Systems* (Institute of Mathematical Statistics, 2008).

[21] F. R. Pinheiro, P. J. van Leeuwen, and G. Geppert, *Efficient nonlinear data assimilation using synchronization in a particle filter,* Quarterly Journal of the Royal Meteorological Society 145, 2510 (2019).

[22] R. Potthast, A. Walter, and A. Rhodin, *A Localized Adaptive Particle Filter within an Operational NWP Framework,* Monthly Weather Review 147, 345 (2018).

[23] M. Ades and P. J. van Leeuwen, *An exploration of the equivalent weights particle filter,* Quarterly Journal of the Royal Meteorological Society 139, 820 (2013).

[24] P. J. van Leeuwen, *Nonlinear data assimilation in geosciences: An extremely efficient particle filter,* Quarterly Journal of the Royal Meteorological Society 136, 1991 (2010).

[25] M. Ades and P. J. van Leeuwen, *The Effect of the Equivalent-Weights Particle Filter on Dynamical Balance in a Primitive Equation Model,* Monthly Weather Review 143, 581 (2014).

[26] M. Ades and P. J. van Leeuwen, *The equivalent-weights particle filter in a high-dimensional system,* Quarterly Journal of the Royal Meteorological Society 141, 484 (2015).

[27] B. R. Hunt, E. J. Kostelich, and I. Szunyogh, *Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter,* Physica D: Nonlinear Phenomena Data Assimilation, 230, 112 (2007).

[28] P. J. van Leeuwen, *Nonlinear Ensemble Data Assimilation for the Ocean,* in *Seminar on Recent Developments in Data Assimilation for Atmosphere and Ocean, ECMWF* (2003).

[29] T. Bengtsson, C. Snyder, and D. Nychka, *Toward a nonlinear ensemble filter for high-dimensional systems,* Journal of Geophysical Research: Atmospheres (2003), 10.1029/2002JD002900@10.1002/(ISSN)2169-8996.SPCTME1.

[30] A. Farchi and M. Bocquet, *Review article: Comparison of local particle filters and new implementations,* Nonlinear Processes in Geophysics 25, 765 (2018).

[31] S. G. Penny and T. Miyoshi, *A local particle filter for high dimensional geophysical systems,* Nonlinear Processes in Geophysics Discussions 2, 1631 (2015).

[32] P. Rebeschini and R. van Handel, *Can local particle filters beat the curse of dimensionality?* The Annals of Applied Probability 25, 2809 (2015).

[33] Y. Lee and A. J. Majda, *State estimation and prediction using clustered particle filters,* Proceedings of the National Academy of Sciences 113, 14609 (2016).

[34] N. Chustagulprom, S. Reich, and M. Reinhardt, *A Hybrid Ensemble Transform Particle Filter for Nonlinear and Spatially Extended Dynamical Systems,* SIAM/ASA Journal on Uncertainty Quantification 4, 592 (2016).

[35] J. Poterjoy, *A Localized Particle Filter for High-Dimensional Nonlinear Systems,* Monthly Weather Review 144, 59 (2016).

[36] J. Poterjoy and J. L. Anderson, *Efficient Assimilation of Simulated Observations in a High-Dimensional Geophysical System Using a Localized Particle Filter,* Monthly Weather Review 144, 2007 (2016).

[37] D. Evans and A. J. Jones, *A proof of the Gamma test,* Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences 458, 2759 (2002).

[38] P. L. Houtekamer and F. Zhang, *Review of the Ensemble Kalman Filter for Atmospheric Data Assimilation,* Monthly Weather Review 144, 4489 (2016).

[39] G. Gaspari and S. E. Cohn, *Construction of correlation functions in two and three dimensions,* Quarterly Journal of the Royal Meteorological Society 125, 723 (1999).

[40] E. N. Lorenz, *Predictability: A problem partly solved,* in *Proc. Seminar on Predictability, Reading, Uk, Ecmwf* (1996).

[41] H. Bal, D. Epema, C. de Laat, R. van Nieuwpoort, J. Romein, F. Seinstra, C. Snoek, and H. Wijshoff, *A Medium-Scale Distributed System for Computer Science Research: Infrastructure for the Long Term,* Computer 49, 54 (2016).

# 4

# A Novel Local Particle Filter Based on Gaussian Process Regression for Highly Nonlinear Observation Operator in High-Dimensional Models

*In data assimilation (DA), various types of observations can be assimilated. Highly nonlinear observation operators are very common in Earth sciences, which goes against the linear and Gaussian assumptions of the Kalman filter. One source of non-linearity that has not been studied widely is the non-linear issue in observation operators. In this research Gaussian process regression (GPR) is used to estimate the uncertainty of non-linear observation operators. At each update, the surrogate is adaptively trained and refined by current observations and model states. When the model states are transferred to the observation space, the surrogate can give the information about estimations and the corresponding uncertainty. A Lorenz model (1996) with 40 variables is used to evaluate the performance of this proposed local particle by conducting a set of experiments with different settings including the number of particles, the impact of localization scales, etc. To test its ability to deal with nonlinear issues, a highly nonlinear observation operator is designed and used in experiments. LETKF and local EAKF are used as benchmarks in this research. The results*

This chapter is based on: Wang, Zhenwu, Rolf Hut, and Nick Van de Giesen. ”A Novel Local Particle Filter Based on Gaussian Process Regression for Highly Nonlinear Observation Operator in High-Dimensional Models” *Journal of Advances in Modeling Earth Systems*(In review).

*show that the new method has a stable performance with high accuracy and it outperforms the two benchmarks. More importantly, for the non-linear case in this study, the new method only uses 25 particles to achieve a good performance. Although only the Lorenz model is considered in this study, it is highly likely that the proposed method will also work with other models.*

## 4.1. Introduction

In data assimilation(DA), the ensemble Kalman filter(EnKF) [1, 2] and its variants are popular and have a broad range of applications. Ensemble-type methods are flexible and easy to implement. They have been successfully applied in Earth science for atmospheric, hydrological and oceanographic dynamical systems [3–8].

EnKF-type methods assume Gaussian distributions, and non-Gaussian error distributions lead to sub-optimal results. However, in Earth science, it is common that lots of dynamics models are nonlinear, especially in high-resolution systems. Nonlinear models produce non-Gaussian error distributions, which can be a reason for the failure of EnKF-type methods [9]. Another source of non-linearity are the observations. Observations can have non-Gaussian error distributions, which happens frequently in ocean and land surface modeling [9, 10]. Unfortunately, true error structures of observations and models are typically not well known. Several studies have reported that sampling frequency and the accuracy of observations could be a source of nonlinearity [11–14]. All these nonlinearities challenge Gaussian assumptions and, consequently, it is highly likely that DA methods fail to track transitions and diverge from the truth of model states [15].

For issues of nonlinearity in DA, particle filters(PFs) provide a potential solution. Unlike Ensemble-type methods, PFs relax all Gaussian assumptions [16, 17]. However, for typical particle filters, the number of particles increases exponentially with the dimension of a model. Otherwise, PFs experience weight degeneracy and filter collapse, which is called the curse of dimensionality [18]. When the dimensionality of a model is high, a large number of particles is needed to avoid filter collapse, which is prohibitively costly in most cases [18–20]. Moreover, in typical PFs, particle samples can become impoverished when the model is deterministic [21].

Applying localization methods in PFs can solve the issue of dimensionality [9, 10, 19, 22, 23]. In the context of PFs, the prior particles do not depend on the covariance of particles. Therefore, the implementation of localization in EnKF does not fit into PFs [22, 24]. Depending on where the analysis process of data assimilation is performed, the localization methods in PFs can be divided into two strategies: state–domain localization and sequential–observation localization [10]. The state–domain approach has been applied successfully in several studies [20, 25–27] and the second localization method is used in Poterjoy [24]'s research. Localization is an effective way to solve difficulties of particle filters in high dimensional systems.

Developing nonlinear data assimilation methods has attracted critical attention in recent decades. van Leeuwen [22] proposed a fully nonlinear particle filter by using the proposal density to ensure equal weights for particles. This approach has been tested in the Lorenz model with 40 variables and even higher dimensions using only 20 particles. An iterative ensemble smoother using an adaptive Gaussian process was developed to solve nonlinear issues in hydrological inverse problems [28, 29]. To tackle the nonlinearity in data assimilation, the kernel-based ensemble Gaussian mixture filtering (EnGMF) was introduced by Anderson and Anderson [30]. In this approach, the Gaussian mixture was used to represent the probability distribution function of model states. It included two update steps: a Kalman filter update and a particle filter update. Therefore, this method is a hybrid of the particle filter and the ensemble Kalman filters. Several schemes based on EnGMF has been designed [31–33]. Lei and Bickel [21] presented a debiasing method called the nonlinear ensemble adjustment filter(NLEAF). NLEAF had been evaluated in the strongly nonlinear Lorenz-96 model with satisfactory results. Bocquet *et al.* [9] reviewed non-Gaussian and nonlinear aspects of data assimilation and compared many methods trying to deal with non-Gaussianity and nonlinearity. Nonlinear models and nonlinear observation operators are the sources of nonlinearity in data assimilation. Current nonlinear solutions are still specific to particular situations and are hard to generalize [9].

In this study, we present a local particle filter using Gaussian process regression(GPR) that has the potential to address issues of nonlinear observation operators in data assimilation frequently encountered in geoscience. The GPR model is used to replace a nonlinear observation operator. Because of GPR's properties, when it transfers model states into observation space, GPR model can provide the estimation with the information of observation uncertainty. The use of GPR gives a solution to the issue of non-linear observation operators. For the curse of dimensionality, sequential–observation localization is a useful tool to overcome this obstacle. The use of these two strategies together is the focus of this research.

The Chapter is organized as follows: Section 4.2 provides a detailed description of our proposed algorithm. The sensitivity experiments and corresponding results are given and discussed in Section 4.3. Conclusions are shown in Section 4.4.

## 4.2. Methodology

### 4.2.1. Particle filters with localization

Chapter 2 introduces the standard particle filters in Section 2.1 and Gaussian Process Regression in Section 2.4. In this section, we will give the implementation of the new particle filters. A localization method ensures that model states are updated by near observations within the localization scale and the influence of distant observations is removed. Localization has been proven to be a useful scheme to prevent filter degeneracy [19, 24, 25, 27].

The influence of the localization on particle filters is achieved by Eq.(4.1).

$$\omega_{n,j} = p\left(y|x_{n,j}\right) l\left(y, x_j, r\right), \begin{cases} j = 1, \ldots, N_x \\ n = 1, \ldots, N_{ens} \end{cases} \tag{4.1}$$

where $l\left(y, x_j, r\right)$ is the localization function and $r$ is the localization radius. The localization function used in the current study has a Gaussian-type structure with a specific radius $r$ [34]. Coefficients calculated by the localization function are determined by the Euclidean distance between observations $\boldsymbol{y}$ and model states $\boldsymbol{x}$. The maximum value of local coefficients is 1 when the distance is 0, and its value decreases to 0 when the distance becomes larger. By multiplying localization coefficients with weights, the influence of local observations is reflected by the localization coefficients. The definition of the distance depends on the specific models and needs a certain prior knowledge of physical length to be determined. We assume that errors are independent in observations. Thus, $p\left(\boldsymbol{y}|\boldsymbol{x}_n\right)$ can be calculated by $\prod_{i=1}^{N_y} p\left(y_i|\boldsymbol{x}_n\right)$. When observations are assimilated one by one, the weights of the $j$th model state given the $i$th observation can be written by:

$$\begin{aligned}\omega_{n,j}^{(y_i)} &= \prod_{i=1}^{q} p\left(y_q|x_{n,j}^{(y_0)}\right) l\left(y_i, x_j, r\right) \\ &= \omega_{n,j}^{(y_{i-1})} p\left(y_q|x_{n,j}^{(y_0)}\right) l\left(y_q, x_j, r\right)\end{aligned} \tag{4.2}$$

$$W_j^{(y_i)} = \frac{\omega_{n,j}^{(y_i)}}{\sum_{n=1}^{N_{ens}} \omega_{n,j}^{(y_i)}} \tag{4.3}$$

where $x_{n,j}^{(y_0)}$ denotes the prior particles before assimilating $y_i$ and $W_j^{(y_i)}$ are the normalized weights. Because the use of localization, the value of $l\left(y_i, x_j, r\right)$ can be 0. If all of the $\omega_{n,j}^{(y_i)}$ is 0, all weights will be set to 1.

### 4.2.2. Local particle filters based on Gaussian process regression

The local PF with GPR(LPF-GPR) is inspired by the local particle filters proposed by Poterjoy [24]. Major steps of the proposed algorithm are shown in this section, and the pseudocode description of this algorithm is given.

In LPF-GPR, all observations are assimilated one by one, just like the algorithm in Poterjoy [24]'s research. For $i$th observation $y_i$ for $i = 1, \ldots, N_y$, before assimilating $y_i$, the prior particles can be denoted $\boldsymbol{x}_n^{y_{i-1}}$ for $n = 1, \ldots, N_{ens}$. Based on Bayesian theory, the posterior particles $\boldsymbol{x}_p^{(y_{i-1})}$ will be created by resampling prior particles with weights $\tilde{w}_n = p\left(y_i|\boldsymbol{x}_n^{(y_{i-1})}\right)$ for each particle, which can be calculated by using Eq.(4.2). The mean and standard deviation of the probability density function used for getting weights $\tilde{w}$ are from a GPR model and are estimated according to Eq.(2.19). $\tilde{W}$ denotes the normalized $\tilde{w}$. The particles with high weights, which are retained by a resampling algorithm, are duplicated and replace particles with lower weights and make the number of posterior particles consistent with the prior's. Thus, the index $p$ is equal to index $n$. In general, the mean of posterior particles denoted as $\overline{\boldsymbol{x}}^{(y_i)}$ is given by sampled particles generated by a resampling algorithm. To make each particle more unique and avoid filter collapse, a merging step is taken to combine the prior particles $\boldsymbol{x}_n^{y_{i-1}}$ and sampled particles $\boldsymbol{x}_p^{(y_{i-1})}$ by using two coefficients vectors $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$. The final updated particles $\boldsymbol{x}_n^{(y_i)}$ are shown in Eq.(4.4).

$$\boldsymbol{x}_n^{(y_i)} = \overline{\boldsymbol{x}}^{(y_i)} + \boldsymbol{r}_1 \circ \beta\left(\boldsymbol{x}_p^{(y_{i-1})} - \overline{\boldsymbol{x}}^{(y_i)}\right) + \boldsymbol{r}_2 \circ (1-\beta) * \left(\boldsymbol{x}_n^{(y_{i-1})} - \overline{\boldsymbol{x}}^{(y_i)}\right) \tag{4.4}$$

where $\circ$ is an element-wise vector product and $\beta$ is a scalar. The detailed derivations of $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$ can be found in the Appendix of Poterjoy [24]'s study, and here we just show how to calculate them in Equations (4.5) to (4.7).

$$r_{1,j} = \sqrt{\frac{\sigma_j^{{(y_i)}^2}}{\frac{1}{N_{ens}-1}\sum_{n=1}^{N_{ens}}\left[x_{p,j}^{(y_{i-1})} - \bar{x}_j^{(y_i)} + c_j\left(x_{n,j}^{(y_{i-1})} - \bar{x}_j^{(y_i)}\right)\right]^2}} \tag{4.5}$$

$$r_{2,j} = c_j r_{1,j} \tag{4.6}$$

$$c_j = \frac{N_{ens}\left(1 - l\left(x_j, y_i, r\right)\right)}{l\left(x_j, y_i, r\right)\tilde{W}} \tag{4.7}$$

where $j = 1, \ldots, N_x$, and $\sigma_j^{{(y_i)}^2}$ is the error covariance conditioned on all observations up to $y_i$. As $l\left(x_j, y_i, r\right)$ approaches 1, the posterior variance is close to the variance of sampled particles. When $l\left(x_j, y_i, r\right) = 1$, $r_{2,j} = 0$ and all weights are put onto the sampled particles, which means the sampled particles are the posterior particles. As $l\left(x_j, y_i, r\right)$ approaches 0, the posterior variance approximates the prior particles'. When $l\left(x_j, y_i, r\right) = 0$, $r_{1,j} = 0$ and all weight is given to the prior particles, which means particles stay unchanged.

Except for localization, an extra method is used to improve the stability of the proposed method by making weights more uniform. Poterjoy [24]'s local particle filters calculated weights by using:

$$w_n^{y_i} = \left[p\left(y_i|\boldsymbol{x}_n\right) - 1\right] l\left(y_i, \boldsymbol{x}_n, r\right)\alpha + 1 \tag{4.8}$$

where $\alpha$ is a fixed value. The role of Eq.(4.8) is similar to the inflation method, which is commonly used in ensemble-type filters [35]. Next, we introduce a parameter named the mean effective number of ensemble $N_{eff}$ defined as in Eq.(4.9). In general, $N_{eff}$ is used to evaluate the quantity of the ensemble [25]. Here, we applied it to change the inflation factor adaptively.

$$N_{eff} = \left[\sum_{n=1}^{N_{ens}}\left(w_n\right)^2\right]^{-1} \tag{4.9}$$

Typically, weights are calculated based on the Gaussian probability density function (PDF) directly. In this case, the collapse of particle filters happens often. Eq.(4.8) provides a possible solution to avoid collapse because the differences between weights are narrowed down and more particles are kept after the resampling step. This strategy can prevent filter degeneracy and stabilizes filters effectively. In this study, we attempted to calculate weights in an adaptive way by using:

$$w_n^{y_i} = \left[p\left(y_i|\boldsymbol{x}_n\right) l\left(y_i, \boldsymbol{x}_n, r\right) + \alpha_1\right] * \alpha_2 \tag{4.10}$$

The coefficients $\alpha_1$ and $\alpha_2$ can be tuned to make $N_{eff}$ approach a certain value using a bisection algorithm. To achieve this, $\alpha_1$ and $\alpha_2$ are changed dynamically, and the search for these two parameters stops when the first pair of $\alpha_1$ and $\alpha_2$ is found. The reason why we chose the form of Eq.(4.10) is that this form provides a wider range of the value of $w_n^{y_i}$. More effective ways to calculate weights will be investigated in future studies. Consequently, Eq.(4.2) is rewritten into:

$$\begin{aligned} \omega_{n,j}^{(y_i)} &= \prod_{i=1}^{N_y}\left[p\left(y_i|x_{n,j}^{(y_0)}\right) l\left(y_i, x_j, r\right) + \alpha_1\right] * \alpha_2 \\ &= \omega_{n,j}^{(y_{i-1})}\left[p\left(y_i|x_{n,j}^{(y_0)}\right) l\left(y_i, x_j, r\right) + \alpha_1\right] * \alpha_2 \end{aligned} \tag{4.11}$$

The pseudocode description of LPF-GPR is given in Algorithm 2. Although this algorithm is derived from Poterjoy [24]'s local particle filters, it still has some major differences with that study, which are listed here. First, the observation operators are replaced with GPR models to cope with nonlinearity, and the error distribution in those observation is provided by GPR models. Therefore, the way to calculate weights is different. Next, after processing all observations, Poterjoy [24] used a probability mapping method named kernel density distribution mapping (KDDM) to map prior particles to match desired posterior particles. In our algorithm, the assimilation procedure is finished when particles are updated without extra adjustment or correction. Last, as described in Poterjoy [24]'s study, the fixed inflation factor was used to prevent the filter collapse. In our method, we tried to keep $N_{eff}$ close to a certain value by reweighting particles with flexible inflation factors.

## 4.3. Numerical experiments and results

### 4.3.1. The Lorenz(1996) model

The Lorenz 1996 model(Lorenz 96) [36] is a low-order, discrete, chaotic model, which evolves in time according to the following set of differential equations:

$$\frac{\mathrm{d}x_n}{\mathrm{d}t} = (x_{n+1} - x_{n-2})\, x_{n-1} - x_n + F, \quad n = 1 \ldots N_\mathrm{m} \tag{4.12}$$

where $x_n$ represents a vector with state variables, $x_n$ for $n = 1, \ldots N_m$ with periodic boundary conditions: $x_{n+N_m} = x_n$ and $x_{n-N_m} = x_n$, and $N_\mathrm{m}$ is the dimension of the system. The differential equations in Equation (4.12) are integrated by a fourth-order Runge–Kutta method. The time step in the integration is usually set to 0.05, which represents 6 hours of real-time. The value of $F$ can determine the degree of chaos in the Lorenz system, and the choice of $N_m$ can take arbitrary values. In this research, we used typical configurations of $F$ and $N_m$, which were set to 8 and 40, respectively, leading to chaotic behavior.

### 4.3.2. Experiment settings

A series of data assimilation experiments were conducted to evaluate the performance of LPF-GPR. These tests provide insight into the sensitivity to different settings, including the number of particles and localization scales, etc., which are useful for applying LPF-GPR in a real application. LPF-GPR uses a Gaspari and Cohn [34] correlation function for the implementation of localization. In this study, we assume that no model errors exist in the system, so when propagating the model, no model errors are added to the whole system. The observations are from the truth with an uncertainty

**Algorithm 2** A pseudocode description of the LPF-GPR

```
for $i = 1 \rightarrow N_y$ do
    for $n = 1 \rightarrow N_{ens}$ do
        $p\left(y_i|\boldsymbol{x}_n^{(y_{i-1})}\right) \leftarrow GPR\left(y_i, \boldsymbol{x}_n^{(y_{i-1})}\right)$
        $\tilde{w}_n \leftarrow \left[p\left(y_i|\boldsymbol{x}_n^{(y_{i-1})}\right) + \alpha_1\right]\alpha_2$
        (Tuning $\alpha_1$ and $\alpha_2$ is to keep $N_{eff}$ fixed. $N_{eff} = \left[\sum_{n=1}^{N_{ens}} (\tilde{w}_n)^2\right]^{-1}$)
    end for
    $\tilde{W} \leftarrow \sum_{n=1}^{N_e} \tilde{w}_n$
    $\tilde{W_n} \leftarrow \tilde{w}_n/\tilde{W}$
    Generate new particles $\boldsymbol{x}_p^{(y_{i-1})}$ based on normalized $\tilde{W_n}$ using the residual resampling algorithm.
    for $j = 1 \rightarrow N_x$ do
        for $n = 1 \rightarrow N_{ens}$ do
            $p\left(y_i|x_{n,j}^{(y_0)}\right) \leftarrow GPR\left(y_i, x_{n,j}^{(y_0)}\right)$
            $\omega_{n,j}^{(y_i)} \leftarrow \omega_{n,j}^{(y_{i-1})}\left[p\left(y_i|x_{n,j}^{(y_0)}\right) l\left(y_i, x_j, r\right) + \alpha_1\right]\alpha_2$
            (Tuning $\alpha_1$ and $\alpha_2$ is to keep $N_{eff}$ fixed. $N_{eff} = \left[\sum_{n=1}^{N_{ens}} (\tilde{w}_n)^2\right]^{-1}$)
        end for
        $\boldsymbol{\Omega}_j^{(y_i)} \leftarrow \frac{\omega_{n,j}^{(y_i)}}{\sum_{n=1}^{N_{ens}} \omega_{n,j}^{(y_i)}}$
        $\bar{x}_j^{(y_i)} \leftarrow \sum_{n=1}^{N_{ens}} \boldsymbol{\Omega}_j^{(y_i)} x_{n,j}^{(y_0)}$
        $\sigma_j^{(y_i)2} \leftarrow \boldsymbol{\Omega}_j^{(y_i)}\left[x_{n,j}^{(y_0)} - \bar{x}_j^{(y_i)}\right]^2$
        $c_j \leftarrow \frac{N_{ens}\left(1-l\left[x_j, y_i, r\right]\right)}{l\left[x_j, y_i, r\right]\tilde{W}}$
        $r_{1,j} \leftarrow \sqrt{\frac{\sigma_j^{(y_i)2}}{\frac{1}{N_{ens}-1}\sum_{n=1}^{N_{ens}}\left[x_{k_{n,j}}^{(y_{i-1})} - \bar{x}_j^{(y_i)} + c_j\left(x_{n,j}^{(y_{i-1})} - \bar{x}_j^{(y_i)}\right)\right]^2}}$
        $r_{2,j} \leftarrow c_j * r_{1,j}$
        $x_{n,j}^{(y_i)} \leftarrow \bar{x}_j^{(y_i)} + \beta * r_{1,j}\left(x_{p,j}^{(y_{i-1})} - \bar{x}_j^{(y_i)}\right) + (1-\beta) * r_{2,j}\left(x_{n,j}^{(y_{i-1})} - \bar{x}_j^{(y_i)}\right)$
    end for
end for
```

$\epsilon \sim N(0, 0.1)$, and half of all state variables are observed with fixed locations, which means 20 observations are assimilated in each experiment. In each test, the first 1000 cycles are a spin-up time. After this corresponding metrics are collected and calculated from 9000 cycles to verify the performance of LPF-GPR. LETKF [37] was used as a benchmark to compare with the performance of LPF-GPR. In addition to localization, a typical inflation method was applied in LETKF to prevent filter collapse during data assimilation.

To make the experiments more challenging, we used non-linear observations generated from the truth using a non-linear H operator. The specific form of the non-linear H is shown below, which is the default setting for all experiments.

$$\boldsymbol{y} = 0.2 * \log(|\boldsymbol{x}|) - 0.01 * e^{\boldsymbol{x}/10} + 1.5 * \boldsymbol{x} - 2.5 * \sqrt{|\boldsymbol{x}|} + 0.2 * (\boldsymbol{x}/5)^2 \tag{4.13}$$

The root mean square error(RMSE) and ensemble spread are used to evaluate the proposed method. The ensemble spread is defined as the square-root of the ensemble variance averaged over all model states. The definitions of these two metrics are as follows.

$$\text{RMSE} = \sqrt{\frac{1}{N_y}\sum_{k=1}^{N_y}\left(x_{\text{truth},k} - x_k^t\right)^2} \tag{4.14}$$

$$\text{Spread} = \sqrt{\frac{1}{N_y}\sum_{k=1}^{N_y}\sum_{i=1}^{N_{\text{ens}}}\left(x_{i,k}^t - x_k^t\right)^2} \tag{4.15}$$

### 4.3.3. Results

To evaluate the stability of LPF-GPR using a small number of particles, the first experiment uses a scenario in which the default configuration was used. The current posterior particles are the prior particles for the next cycle. We used LETKF as a benchmark in the tests. Prior error statistics calculated from the domain averages of RMSE and spread of all time steps, and all model states were collected during data assimilation for both filters. Figure 4.1 shows the time series of the second model state, which was unobserved and domain-averaged prior RMSE and spread over time. The plot started at the beginning of the data assimilation, and the new filter took some time to reach a stable status. The averaged RMSE of LETKF was much higher than LPF-GPR, and its spread decreases rapidly. LETKF indeed suffers from filter collapse. The non-linear operator leads to suboptimal results for LETKF, which is the reason for the filter divergence. The prior statistics show that The LPF-GPR provides a stable result with satisfying accuracy using a small number of particles. The RMSE and spread averaged over the particles for each cycle are 0.37 and 0.71, respectively. However, forecast errors are higher than the observation errors, possibly caused by sampling errors and non-linearity in observations. This experiment demonstrates that LPF-GPR outperforms LETKF and can give stable results using a small number of particles.

Next, we explore the impact of parameters $\alpha 1$, $\alpha 2$ and $\beta$ in the new filter on the performance of LPF-GPR when using different localization scales. As mentioned in the

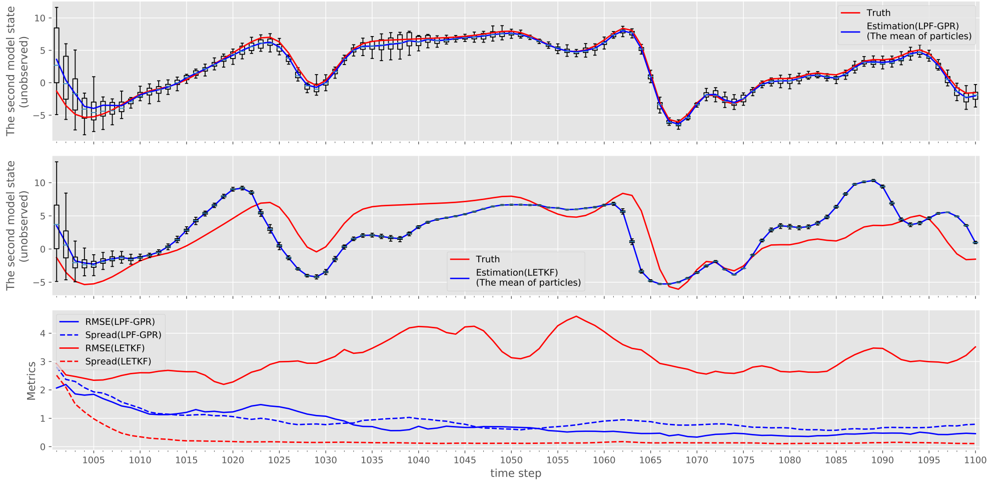

Figure 4.1: Trajectories of (top) the second model state which is unobserved using the LPF-GPR, ( middle) the second model state using the LETKF and (bottom) RMSE and spread for both methods. In the first two figures, the blue curve is the truth and the red curve is the estimation of the model state. In the third one, the solid lines indicate the RMSE and the dash lines is the spread. The assimilation starts at time step 1,000.

last section, parameter $\beta$ comes from Eq.(4.4) and $\alpha$ comes from in Eq.(4.8). The localization procedure is of great importance because, with its use, the collapse of particle filters is avoided successfully. The choice of the localization scale imposes a significant impact on the performance of LPF-GPR, which can determine the number of observations assimilated and the efficiency of data assimilation. Consequently, we explore a part of the parameter space of these three parameters. As mentioned before, the number of possible combinations of $\alpha 1$ and $\alpha 2$ is huge. But the search for $\alpha 1$ and $\alpha 2$ stops when the first pair is found. To search more efficiently, we make $\alpha 1$ and $\alpha 2$ have the same value.

Figure 4.2 compares prior mean RMSEs and spread from a series of experiments when the proposed filter is applied for various combinations of these three parameters. When the localization scale is increased to 5, LPF-GPR yields the lowest RMSE. Nevertheless, the continuous increase in the localization scale does not achieve better results, possibly because more observations are more likely to obtain a higher number of degrees of freedom, impacting the calculation of weights. Results from experiments applying different beta do not show significant changes, which means LPF-GPR responds similarly when $\beta$ is changed. These experiments provide a guide for choosing the proper localization scale, $\beta$ and $\alpha$ with the current configuration. It also suggests that the LPF-GPR exhibits more sensitivity to the localization scale than $\beta$ and $\alpha$, which is useful and practical for future applications.

Determining the number of particles used in a data assimilation system is problematic for real applications in Earth science. The curse of dimensionality in particle filters always requires more and more particles, but limited computing resources are the bottleneck for using a large number of particles. The cost of using many particles in large-scale models is prohibitive. For this practical reason, we explore the potential of

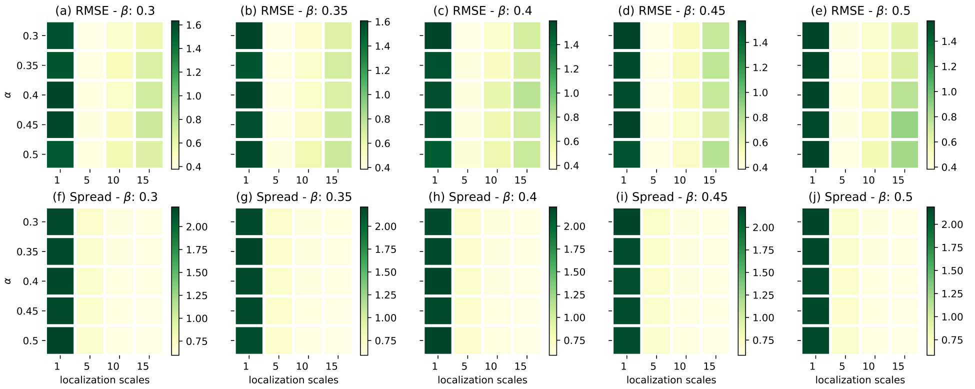

Figure 4.2: Prior mean RMSE(a-e) and spread(f-j) as a function of two parameters $\alpha$ and $\beta$ when using localization scales.

the new filter when using as few as possible particles. Results from those experiments are shown in Figure 4.3.

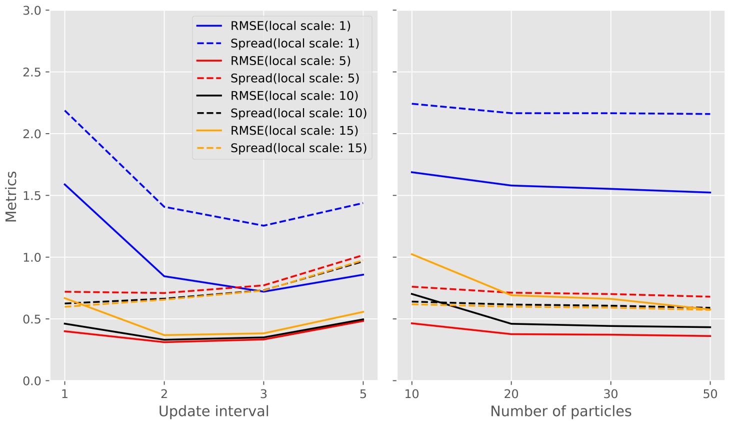

Figure 4.3: Prior mean RMSE and spread as a function of (left)assimilation interval and (right) number of particles for experiments with different localization scales.

LPF-GPR achieves a satisfactory performance using a small number of particles, which is a practical benefit for real applications. For LPF-GPR, ten particles can provide stable results with relatively high accuracy. In general, more particles can help particle filters overcome the curse of dimensionality and improve its accuracy. However, based on results in Figure 4.3, using more particles does not improve the performance significantly. There are two possible explanations for these results. First, the results indicate that the localization procedure has prevented the filter collapse successfully for fewer particles. In the meantime, the use of localization can break the consistency of Bayes theorem in particle filters. The imbalance caused by localization stops the increase in accuracy by using more particles. Another possible reason is that the Gaussian process regression can capture uncertainty in this highly non-linear case effectively. More particles do not provide a more accurate approximation for the distribution of state variables. Besides, LPF-GPR is less sensitive to the change of localization scales. In Figure 4.3, it is clear that the optimal localization scale is five, and the filter performs worst in the case with localization scale one. One of the main objectives of using localization is to reduce the impact of sampling errors from particles. Thus, it is highly likely that the increasing localization scale can bring more sampling errors. The results in Figure 4.3 show that when using larger localization scales in experiments, the filter performance decreases slightly. One possible reason for this is that the filter can deal with errors that may happen in the process of data assimilation.

To assess the proposed filter performance further, several experiments were conducted when observations were assimilated at different frequencies. A time step of 0.05 units in the Lorenz model represents six hours. Therefore, time interval 1, 2 3 and 5 mentioned in Figure 4.3 are 6, 12, 18 and 60 hours respectively. The number of data assimilation cycles is set to the default value, which means model states are updated less frequently when time interval of assimilation is increased.

Increasing update interval brings more nonlinearity to data assimilation because of changes in model errors [21, 25]. In general, model errors are accumulated with model propagation over time when assimilating observations at lower frequencies. Consequently, it leads to worse estimations of model states.

Figure 4.3 shows that data assimilation with more intervals, achieves a better performance. The nonlinearity is most likely the reason for this. The nonlinear observation operator leads to the non-Gaussian distribution of observation errors, and it brings more non-Gaussianity to forecast errors in assimilation cycles than the purely linear case. It is highly likely that the non-Gaussianity in data assimilation causes changes in Figure 4.3.

To examine the potential impact brought by the nonlinearity in observations on prior particles in data assimilation cycles, we applied the Kolmogorov-Smirnov(KS) test to prior particles to determine whether a significant deviation from normality exists in it. Particles of the first two states $x_1$ and $x_2$ were examined by the KS test in each assimilation cycle over time. Figure 4.4 shows the percentage of cases in which the prior particles fail the KS test at the 5% significance level for each experiment.

As illustrated in this figure, increasing the update interval can reduce the percentage of non-Gaussianity for both model states. It is possibly caused by the accumulation of forecast errors in the model propagation over time. When we assimilate observations less frequently, forecast errors are close to the climatological errors, which follows a Gaussian distribution approximately in the L96 model [24]. Lower assimilation frequencies allow model errors to accumulate over time and increase the Gaussianity in it. We can conclude that, from Figure 4.4, more Gaussianity in model errors can improve the performance of LPF-GPR to some extent. But when the update interval of the observation network is set to five, the corresponding metrics show that the filter does not improve the accuracy of results. One possible explanation for this result is that, although the distribution of model errors approximates the Gaussian with model propagation, in the meantime, it increases the accumulation of model errors. In this case, model errors have a more dominant and negative impact on the accuracy of results than the benefits of the Gaussianity.

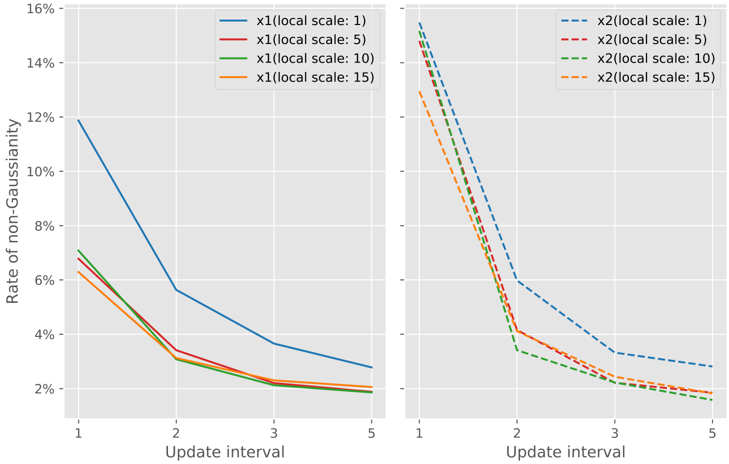

Figure 4.4: Percentage of cycles for the first two states that fails the KS (Kolmogorov-Smirnov) test at the 5% significance level in experiments when using different assimilation intervals and localization scales.

The information provided by the RMSE and the spread of particles for the filter's probabilistic skill is limited. Therefore, we attempted to use rank histograms to investigate the phenomenon observed in Figure 4.5. The rank histogram is a verification tool to examine some qualities of particles. When plotting rank histograms, all particles of a state are ranked in increasing order. If the truth of a model state falls in intervals formed by ranking particles in ascending order with equal probability, it means the truth is statistically indistinguishable from particles, and all particles are from the same distribution [38]. It can be an indicator of the reliability of data assimilation. Therefore, the flatness of rank histograms is a necessary condition but not insufficient [39]. In Figure Figure 4.5, data used to plot rank histograms are from experiments when localization scale five is used, and rank histograms are calculated for every eighth model state using 20 particles using different update intervals.

As shown in Figure 4.5, when assimilating observations at every timestep, the nonuniform rank histograms are observed more frequently in chosen state variables. It suggests deficiencies in probabilistic analysis in this case. In cases with observation networks that measure states less often, rank histograms are relatively uniform, which means that results produced in these experiments are reliable. From the results of rank histograms, we can still notice the influence of model errors in different ways. The Gaussianity provided by model errors, which is good for the performance, has a more significant role when observations are assimilated at proper frequencies. Ranks histograms in the case with less Gaussianity (the first row in Figure 4.5 ), are relatively nonuniform. It seemed that when model errors have an adverse impact on the accuracy, rank histograms, in this case, do not show a strong nonuniformity.

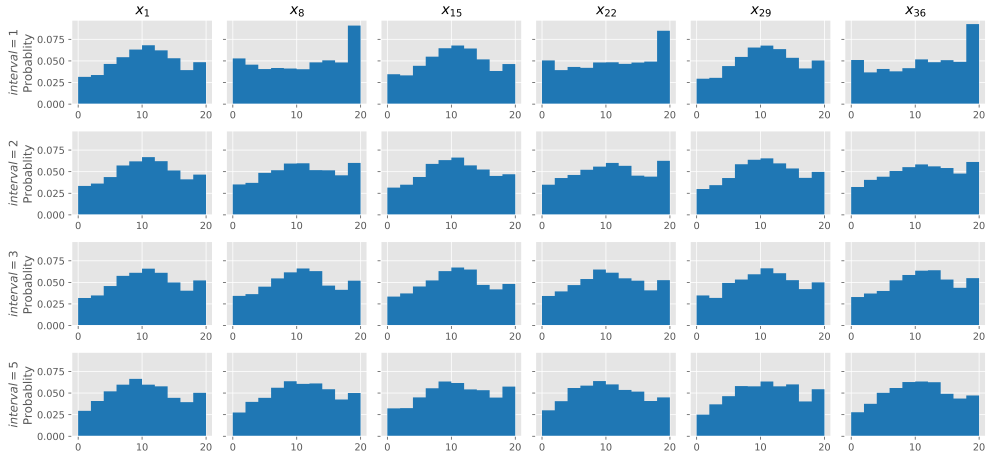

Figure 4.5: Rank histograms for state variable 1, 8, 15, 22, 29 and 33 calculated from experiments with 20 particles when using different assimilation intervals.

## 4.4. Conclusions

In this paper, we proposed a new local particle filter with Gaussian process regression for highly nonlinear observation operators in a high-dimensional Lorenz model. The analysis step of the proposed method was performed on each observation site, and all particles were updated after observations were assimilated. Similar to the localization strategy in Ensemble-type methods, local particle filters only used observations within a localization scale. The influence of distant observations was removed by using a localization function, and the use of localization in PFs avoids filter collapse successfully. Another critical point of this research was to use Gaussian process regression surrogate models to account for nonlinearity in a highly nonlinear observation operator. Because of the properties of the GPR surrogates, uncertainty in observation operators can be quantified, which can be used to calculate weights of particles. Besides, we found an effective way to prevent filters from collapsing. Before applying the resampling algorithm to obtain posterior particles, adjusting the values of weights by tuning the effective number of particles can avoid weight degeneracy.

The localized PFs method with the GPR surrogate presented in this study may be problematic for some specific geoscience applications. It is possible that the use of localization can break the physical consistency in model space. The imbalance brought by localization has been observed in both EnKF [40] and PFs [24]. Previous results in Section 4.3 have shown that the imbalance exists when using different localization scales. Therefore, tuning the localization radius is an effective way to improve the performance of DA. But for some large-scale applications, tuning parameters is impossible due to limited computational resources. For this reason, LPF-GPR shows substantial benefits. From the results in the preceding section, although changes of localization can influence LPF-GPR, it is not sensitive to larger localization scales, which is a practical advantage when implementing LPF-GPR in real cases.

The uncertainty of observations becomes complex and unquantified because of the existing nonlinearity in observations. Constructing a surrogate model to solve nonlinearity in data assimilation is feasible, and it provides a way to make uncertainty in the nonlinear case approximate the Gaussian. Introducing GPR models can bring more uncertainty from a different source, but this method has the ability to deal with uncertainty in data assimilation. It should be noted that, although replacing nonlinear observation operators with GPR surrogates can estimate its uncertainty well, training and using GPR surrogates to calculate corresponding mean and standard deviation takes more computing time.

LPF-GPR has been evaluated by a Lorenz model with 40 variables, and results from experiments showed that, when using highly nonlinear observation operators, the new proposed local PFs only needed 20 particles to avoid weights degeneracy, but the benchmark LETKF with inflation experienced filter collapsed. In this nonlinear case, LPF-GPR has considerable advantages over LETKF, and using a small number of particles is another practical benefit for LPF-GPR. These promising results give a possibility to explore its potential in high-dimensional geophysical applications with highly nonlinear measurement operators. A future topic will focus on investigating its potential benefits and identifying relative weaknesses of LPF-GPR in data assimilation within a high-dimensional geophysical systems.

## References

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

[2] G. Evensen, *The Ensemble Kalman Filter: Theoretical formulation and practical implementation,* Ocean Dynamics 53, 343 (2003).

[3] Y. Zhang, J. Hou, J. Gu, C. Huang, and X. Li, *SWAT-Based Hydrological Data Assimilation System (SWAT-HDAS): Description and Case Application to River Basin-Scale Hydrological Predictions,* Journal of Advances in Modeling Earth Systems 9, 2863 (2017).

[4] R. Abolafia-Rosenzweig, B. Livneh, E. Small, and S. Kumar, *Soil Moisture Data Assimilation to Estimate Irrigation Water Use,* Journal of Advances in Modeling Earth Systems 11, 3670 (2019).

[5] C. Irrgang, J. Saynisch, and M. Thomas, *Utilizing oceanic electromagnetic induction to constrain an ocean general circulation model: A data assimilation twin experiment,* Journal of Advances in Modeling Earth Systems 9, 1703 (2017).

[6] J. Jin, A. Segers, A. Heemink, M. Yoshida, W. Han, and H.-X. Lin, *Dust Emission Inversion Using Himawari-8 AODs Over East Asia: An Extreme Dust Event in May 2017,* Journal of Advances in Modeling Earth Systems 11, 446 (2019).

[7] J. Jin, H. X. Lin, A. Segers, Y. Xie, and A. Heemink, *Machine learning for observation bias correction with application to dust storm data assimilation,* Atmospheric Chemistry and Physics 19, 10009 (2019).

[8] A. M. Fox, T. J. Hoar, J. L. Anderson, A. F. Arellano, W. K. Smith, M. E. Litvak, N. MacBean, D. S. Schimel, and D. J. P. Moore, *Evaluation of a Data Assimilation System for Land Surface Models Using CLM4.5,* Journal of Advances in Modeling Earth Systems 10, 2471 (2018).

[9] M. Bocquet, C. A. Pires, and L. Wu, *Beyond Gaussian Statistical Modeling in Geophysical Data Assimilation,* Monthly Weather Review 138, 2997 (2010).

[10] A. Farchi and M. Bocquet, *Review article: Comparison of local particle filters and new implementations,* Nonlinear Processes in Geophysics 25, 765 (2018).

[11] P. J. van Leeuwen, *Nonlinear Data Assimilation for high-dimensional systems,* in *Nonlinear Data Assimilation*, Frontiers in Applied Dynamical Systems: Reviews and Tutorials, edited by P. J. Van Leeuwen, Y. Cheng, and S. Reich (Springer, 2015) pp. 1–73.

[12] J. L. Anderson, *A Non-Gaussian Ensemble Filter Update for Data Assimilation,* Monthly Weather Review 138, 4186 (2010).

[13] R. N. Miller, M. Ghil, and F. Gauthiez, *Advanced Data Assimilation in Strongly Nonlinear Dynamical Systems,* Journal of the Atmospheric Sciences 51, 1037 (1994).

[14] R. N. Miller, E. F. Carter, and S. T. Blue, *Data assimilation into nonlinear stochastic models,* Tellus A: Dynamic Meteorology and Oceanography 51, 167 (1999).

[15] M. Verlaan and A. W. Heemink, *Nonlinearity in Data Assimilation Applications: A Practical Method for Analysis,* Monthly Weather Review 129, 1578 (2001).

[16] P. J. van Leeuwen, *A Variance-Minimizing Filter for Large-Scale Applications,* Monthly Weather Review 131, 2071 (2003).

[17] N. J. Gordon, D. J. Salmond, and A. F. M. Smith, *Novel approach to nonlinear/non-Gaussian Bayesian state estimation,* IEE Proceedings F (Radar and Signal Processing) 140, 107 (1993).

[18] T. Bengtsson, P. Bickel, and B. Li, *Curse-of-Dimensionality Revisited: Collapse of the Particle Filter in Very Large Scale Systems* (Institute of Mathematical Statistics, 2008).

[19] C. Snyder, T. Bengtsson, P. Bickel, and J. Anderson, *Obstacles to High-Dimensional Particle Filtering,* Monthly Weather Review 136, 4629 (2008).

[20] P. Rebeschini and R. van Handel, *Can local particle filters beat the curse of dimensionality?* The Annals of Applied Probability 25, 2809 (2015).

[21] J. Lei and P. Bickel, *A Moment Matching Ensemble Filter for Nonlinear Non-Gaussian Data Assimilation,* Monthly Weather Review 139, 3964 (2011).

[22] P. J. van Leeuwen, *Particle Filtering in Geophysical Systems,* Monthly Weather Review 137, 4089 (2009).

[23] S. Vetra-Carvalho, P. J. van Leeuwen, L. Nerger, A. Barth, M. U. Altaf, P. Brasseur, P. Kirchgessner, and J.-M. Beckers, *State-of-the-art stochastic data assimilation methods for high-dimensional non-Gaussian problems,* Tellus A: Dynamic Meteorology and Oceanography 70, 1 (2018).

[24] J. Poterjoy, *A Localized Particle Filter for High-Dimensional Nonlinear Systems,* Monthly Weather Review 144, 59 (2016).

[25] S. G. Penny and T. Miyoshi, *A local particle filter for high dimensional geophysical systems,* Nonlinear Processes in Geophysics Discussions 2, 1631 (2015).

[26] N. Chustagulprom, S. Reich, and M. Reinhardt, *A Hybrid Ensemble Transform Particle Filter for Nonlinear and Spatially Extended Dynamical Systems,* SIAM/ASA Journal on Uncertainty Quantification 4, 592 (2016).

[27] Z. Wang, R. Hut, and N. Van de Giesen, *A Local Particle Filter Using Gamma Test Theory for High-Dimensional State Spaces,* Journal of Advances in Modeling Earth Systems 12, e2020MS002130 (2020).

[28] J. Zhang, W. Li, L. Zeng, and L. Wu, *An adaptive Gaussian process-based method for efficient Bayesian experimental design in groundwater contaminant source identification problems,* Water Resources Research 52, 5971 (2016).

[29] L. Ju, J. Zhang, L. Meng, L. Wu, and L. Zeng, *An adaptive Gaussian process-based iterative ensemble smoother for data assimilation,* Advances in Water Resources 115, 125 (2018).

[30] J. L. Anderson and S. L. Anderson, *A Monte Carlo Implementation of the Nonlinear Filtering Problem to Produce Ensemble Assimilations and Forecasts,* Monthly Weather Review 127, 2741 (1999).

[31] I. Hoteit, D.-T. Pham, G. Triantafyllou, and G. Korres, *A New Approximate Solution of the Optimal Nonlinear Filter for Data Assimilation in Meteorology and Oceanography,* Monthly Weather Review 136, 317 (2008).

[32] I. Hoteit, X. Luo, and D.-T. Pham, *Particle Kalman Filtering: A Nonlinear Bayesian Framework for Ensemble Kalman Filters,* Monthly Weather Review 140, 528 (2011).

[33] B. Liu, B. Ait-El-Fquih, and I. Hoteit, *Efficient Kernel-Based Ensemble Gaussian Mixture Filtering,* Monthly Weather Review 144, 781 (2016).

[34] G. Gaspari and S. E. Cohn, *Construction of correlation functions in two and three dimensions,* Quarterly Journal of the Royal Meteorological Society 125, 723 (1999).

[35] J. S. Whitaker and T. M. Hamill, *Evaluating Methods to Account for System Errors in Ensemble Data Assimilation,* Monthly Weather Review 140, 3078 (2012).

[36] E. N. Lorenz and K. A. Emanuel, *Optimal Sites for Supplementary Weather Observations: Simulation with a Small Model,* Journal of the Atmospheric Sciences 55, 399 (1998).

[37] B. R. Hunt, E. J. Kostelich, and I. Szunyogh, *Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter,* Physica D: Nonlinear Phenomena Data Assimilation, 230, 112 (2007).

[38] J. L. Anderson, *A Method for Producing and Evaluating Probabilistic Forecasts from Ensemble Model Integrations,* Journal of Climate 9, 1518 (1996).

[39] T. M. Hamill, *Interpretation of Rank Histograms for Verifying Ensemble Forecasts,* Monthly Weather Review 129, 550 (2001).

[40] H. L. Mitchell, P. L. Houtekamer, and G. Pellerin, *Ensemble Size, Balance, and Model-Error Representation in an Ensemble Kalman Filter,* Monthly Weather Review 130, 2791 (2002).

# 5

# Data assimilation of SMAP soil moisture into the PCR-GLOBWB hydrological model to improve discharge estimates via A Novel Local Particle Filter

*Assimilating surface soil moisture data into hydrological models has been shown to improve the accuracy of hydrological model estimations and predictions. For data assimilation applications in hydrology, the ensemble Kalm filter(EnKF) is the most commonly used data assimilation (DA) method. Particle filters are a type of non-Gaussian filter that does not need the normality assumption that the EnKF needs. Adding localization overcomes the curse of dimensionality, which is a problem in normal particle filters. In the present study, we investigated our adaption of the local particle filter based on the Gamma test theory(LPF-GT) to improve discharge estimates by assimilating SMAP satellite soil moisture into the PCR-GLOBWB hydrological model. The study area is the Rhine river basin, driven by forcing data from April 2015 to December 2016. The improved discharge estimates are obtained by using DA to adjust the surface soil moisture in the model. The influence of DA to discharge is not direct but works through the dynamics of the hydrological model. To explore the potential of LPF-GT, several sensitivity experiments were conducted to figure out the impact of localiza-*

Wang, Z., Hut, R., Tangdamrongsub, N., &Van de Giesen, N. (2021) Data assimilation of SMAP soil moisture into the PCR-GLOBWB hydrological model to improve discharge estimates via A Novel Local Particle Filter. *Hydrology and Earth System Sciences* (Submitted).

*tion scales and the number of particles on DA performance. The DA estimates were validated against in situ discharge measurements from gauge stations. To demonstrate the benefits of LPF-GT, EnKF was used as a benchmark in this research. Increases in Nash-Sutcliffe coefficients (0.05%– 38%) and decreases in normalized RMSE (0.02%–3.4%) validated the capability of LPF-GT. Results showed that the impact of localization scale was substantial. The optimal value of the localization scale was obtained by tuning. LPF-GT achieved a satisfactory performance when only using a few particles, even as few as five particles. The sample errors posed an adverse impact on the open-loop results. Further improvement could be achieved by considering reducing sample errors caused by a small number of particles.*

## 5.1. Introduction

Data assimilation is a collection of methods that aim to find the optimal state of a system given both knowledge of the system (a model) as well as observations related to the state. In hydrology, a typical application for data assimilation is to assimilate satellite soil moisture observations into model predictions to enhance hydrological state estimation and thus improve hydrological predictions [1–4]. The currently most commonly used method of data assimilation in hydrology is the Ensemble Kalman Filter (EnKF), a method that heavily relies on the assumption that most errors are Gaussian in nature. Recently, a variant of the Particle Filter [5, 6], the Local Particle Filter with Gamma Test Theory (LPF-GT) with was introduced as a novel method for doing Data Assimilation with the potential to be useful for high dimensional models used in hydrology [7]. In this article we will test LPF-GT for the first time on a distributed hydrological model and compare its performance against an EnKF.

Soil moisture plays a vital role in hydrological processes, as it controls the exchange of water fluxes between the land surface and the atmosphere [3, 8–10]. Hydrological models can be used to simulate various water storage and flux components with reasonable accuracy. However, achieving accurate soil moisture and discharge estimates are challenging due to the unavoidable simplified modeling of complex hydrological processes and uncertainty in the model's parameters and forcing data [2, 11].

Satellite observations can provide information on large scales at given time intervals. There are varied sources of observed soil moisture retrieved from different satellites that are used for data assimilation applications, such as the Soil Moisture and Ocean Salinity mission (SMOS) [12], NASA's Soil Moisture Active Passive (SMAP) [13], the Advanced Microwave Scanning Radiometer, AMSR-E [14], and the Advanced Scatterometer (ASCAT) [15]. Their main advantage over in-situ soil moisture measurements is that they cover complete areas while in situ observations are relatively sparse. Data assimilation can be used to combine various types of satellite observations at different temporal and spatial resolutions with model estimates to improve the water storage and flux simulations [10, 16–18]. Previous studies applied the assimilation of soil moisture data to improve discharge estimates [1–4, 19, 20].

The ensemble Kalman filter(EnKF) has been used often to assimilate soil moisture observations into hydrological models [21–23]. EnKF is a Monte Carlo method, using a sufficiently large ensemble of model states to approximate the posterior distribution, which is assumed to be Gaussian [24]. The time evolution of mean and covariance is based on the ensemble, and the mean of the ensemble represents the best estimate of the true state of the system. However, the computational cost of EnKF is exceptionally high in high dimensional systems, limiting its usage in complex applications [25]. A small ensemble size is commonly used to reduce the computational effort. However, the limited ensemble size introduces errors in the estimation of the uncertainty of the states by having too few samples (ensemble members) and leads to the rank problem [25], which happens when the number of model states and/or observations is much larger than the ensemble size.

Similar to EnKF, particle filters also use random samples of the state of the model (here called particles) to approximate the uncertainty in the state. Since with particle filters the distribution does not have to be Gaussian, particle filters can handle non-Gaussianity in systems [26], which is an advantage over the EnKF. However, to prevent particle collapse, ie. all particles condensing into the same state, a large number of particles is required, and this increases exponentially with the number of dimensions of considered state variables and observations. This problem is known as the curse of dimensionality [27–29]. Previous studies have been devoted to solving this issue via localization [30–33]. In this study, we focus on the local particle filters with the Gamma test theory (LPF-GT) [7], proposed by us, and apply this method to assimilate soil moisture data into the state of a hydrological model to improve discharge estimates. LPF-GT has several distinguishable benefits compared to ensemble-type algorithms. First, obtaining the inverse of a large matrix is not necessary because each model state is updated independently. Moreover, localization used in LPF-GT overcomes the curse of dimensionality, leading to an additional benefit: it can be implemented in parallel for higher computing efficiency. LPF-GT does not need a large number of particles due to the use of localization. This method massively saves on the amount of computing resources for a DA application in large systems. Finally, the Gamma test theory accounts for the uncertainty brought by data assimilation itself and serves to avoid filter collapse.

In the present research, we aim to test if LPF-GT is a useful data assimilation method in hydrology. We do this by assimilating SMAP soil moisture products with a 9 km resolution into the PCR-GLOWBW 2.0 hydrological model [34] and evaluate the performance of data assimilation by comparing the models estimate of river discharge to observations. Considering the advantages of particle filters over EnKF, and that applications of particle filters in hydrology are rare, it is necessary to attempt to apply particle filters with localization as the data assimilation method in this study. A series of sensitivity experiments are conducted to assess the effects of tunable parameters in LPF-GT, including different localization scales and numbers of particles. Our case study area is the Rhine river basin, where in situ discharge data are used for evaluation. The typical EnKF is used as a benchmark.

This paper is structured as follows. Section 5.2 describes the study area and related forcing and validation data sets used in data assimilation. Descriptions of the PCR-GLOBWB hydrological model, EnKF, and local particle filters are presented in Section 3.2. Settings for numerical experiments are included in this section. Next, we analyze results from data assimilation in Section 5.4. Finally, conclusions and corresponding insights for future studies are given in Section 5.5.

## 5.2. Study area and data sets

### 5.2.1. Study area

The Rhine basin covers 185,000 $\mathrm{km}^2$ and runs over 1320 km from the Alps to the North Sea [35]. The Rhine and its tributaries flow through nine countries, and the largest fraction of the basin is located in Germany. Along its course, the Rhine merges with several major tributaries like the Aare, Neckar, Main, and Moselle. The streamflow at Basel, which belongs to the Rhine's upper part, is dominated by snowmelt and rainfall-runoff from the Alps in summer. Nevertheless, during winter, the stream peaks in the lower parts of the Rhine at Lobith, where it enters the Netherlands, is dominated by rainfall [36, 37]. Across the entire Rhine basin, the mean annual precipitation runs from about 500 mm (Rhine valley) to 2000 mm (Alpine region), and at Lobith, the mean annual discharge is roughly 2200 $\mathrm{m}^3/\mathrm{s}$. Figure 5.1 depicts the Rhine basin.

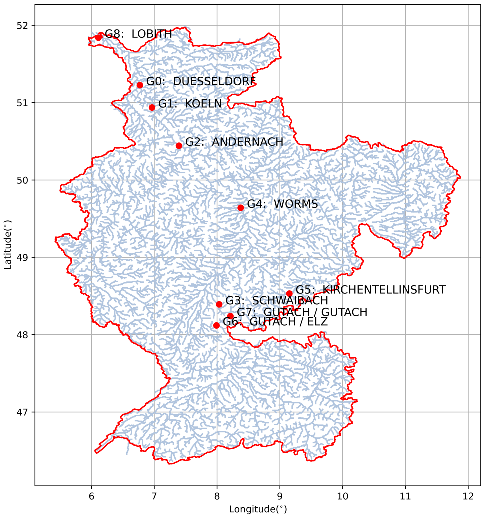

Figure 5.1: Map of the Rhine river basin and water network. Red points indicate locations of river gauge stations used in this paper.

### 5.2.2. SMAP soil moisture

In this study, we used the enhanced level 3 (version 4) soil moisture data product (SPL3SMP_E) retrieved from the NASA Soil Moisture Active Passive (SMAP) radiometer which is distributed on the 9 km Equal-Area Scalable Earth Grid, Version 2.0 (EASE-Grid 2.0) in a global cylindrical projection. The NASA Soil Moisture Active Passive (SMAP) satellite mission was launched on January 31, 2015. It was designed to provide a global high-resolution mapping of soil moisture using an L-band microwave apparatus, which was expected to extend our knowledge of the processes that link the water, energy, and carbon cycles. All data are from the National Snow and Ice Data Center Distributed Active Archive Center (NSIDC DAAC, `https://nsidc.org/data/smap`). Descending SMAP data with retrieval quality flag values of 0 and 8 were used in this study.

### 5.2.3. Forcing data

The forcing data needed by the PCR-GLOWB model are air temperature, precipitation, and reference evaporation (Eref) [34, 40]. Precipitation and air temperature data from 2000 to 2016 were obtained from the European Climate Assessment & Data set and E-OBS gridded dataset (ENSEMBLES project) [38]. Daily Eref data were derived from air temperature data via the Hamon reference evaporation equation [39].

### 5.2.4. Discharge data

The discharge estimates from the model are validated using in situ observations for nine gauges in the Rhine river basin. Daily time series of discharge measurements were acquired from the European Terrestrial Network for River Discharge at the Global Runoff Data Centre (GRDC) (`http://www.bafg.de/GRDC/`). Stations were chosen to represent a large spread in study area. All gauges stations used in this study are shown in Fig. 5.1, and more detailed information on these stations can be found in Table 5.1.

Table 5.1: Gauge stations' information

| No. | Gauge | Longitude | Latitude | River |
|---|---|---|---|---|
| G0 | DUESSELDORF | 6.770183 | 51.225547 | RHINE RIVER |
| G1 | KOELN | 6.963293 | 50.936961 | RHINE RIVER |
| G2 | ANDERNACH | 7.39205 | 50.443386 | RHINE RIVER |
| G3 | SCHWAIBACH | 8.03256 | 48.391719 | KINZIG |
| G4 | WORMS | 8.376019 | 49.64112 | RHINE RIVER |
| G5 | KIRCHENTELLINSFURT | 9.150636 | 48.531054 | NECKAR |
| G6 | GUTACH / ELZ | 7.990062 | 48.119069 | ELZ |
| G7 | GUTACH / GUTACH | 8.212933 | 48.24 | GUTACH |
| G8 | LOBITH | 6.11 | 51.84 | RHINE RIVER |

## 5.3. Methodology

The data assimilation algorithms used in this chapter were introduced in Section 2.2 and Chapter 3, respectively.

### 5.3.1. PCR-GLOBWB hydrological model

PCR-GLOBWB, a global grid-based distributed hydrology model [34, 40], simulates water exchanges between water stocks and fluxes. The implementation of PCR-GLOBWB is based in the PCRaster-Python environment. The spatial resolution of the model is five arcmins ($\sim$10 km × 10 km at the equator) and time steps for all dynamic processes in PCR-GLOBWB are one day. The PCR-GLOBWB model has 3588 gird points in total in the Rhine river basin. The schematic structure of the PCR-GLOBWB model includes five hydrological modules: meteorological forcing, land surface, groundwater, surface water routing, and irrigation and human water use. For each grid cell, PCR-GLOBWB simulates moisture changes in soil layers and water exchange among the soil, the atmosphere, and the groundwater. The generated run-off includes baseflow, surface run-off, interflow, and snowmelt.

Meteorological forcing of PCR-GLOBWB needs time series of precipitation, temperature, and reference evaporation, which can be calculated based on daily mean temperature using the Hamon [39]'s equation or FAO guidelines if other relevant factors are available. There are four types of land cover: paddy irrigated crops, non-paddy irrigated crops, short natural vegetation, and long natural vegetation. Soil and vegetation conditions can be specified for each land cover type. Human water use is included within the hydrological model. Water abstraction and consumptive water use and return flow for irrigation, livestock, industry, and households are considered.

It should be noted that, for this application, PCR-GLOBWB can be set up to without calibration for any given place or globally. All parameters in the model were derived from several geological sources on a global scale. Therefore, each cell grid is associated with multiple and particular parameterizations. The standard parameterizations in PCR-GLOBWB carry land cover, soils, topography, and others, influencing operation schemes for run-off infiltration partitioning, interflow, groundwater recharge, and capillary rise. More detailed information can be found in Sutanudjaja *et al.* [34].

### 5.3.2. Data assimilation setup

Sources of uncertainty in the PCR-GLOBWB model include the meteorological forcing data and model parameter errors. Parameters associated with the top surface soil moisture were perturbed with additive white noise with a standard deviation of 10% of the nominal value. The forcing data error was assumed spatially uncorrelated. The standard deviations of precipitation, air temperature, and reference evaporation were 10%, 15%, 15% of the nominal values, respectively. Propagation of all forcing data and soil moisture in the model was generated by multiplicative Gaussian noise. The observation error is assumed to be 0.04 $m^3/m^3$ based on previous studies [20, 41, 42].

It should be noted that, in the assimilation, after soil moisture was updated by assimilating SMAP observations, each model implementation ran continuously to update underground water and routing submodels in the PCR-GLOBWB model. Consequently, we obtained an ensemble of discharges. The updated discharge estimates were obtained by averaging the discharge ensemble. This approach has been applied in several studies [3, 10]. This procedure was also followed for the estimates of discharge given by the open loop (ie. no data assimilation) runs. For LPF-GT settings, the $N_{eff}$ defined in Equation 14 was set to 0.6 as an optimal value to stabilize the DA process. Similarly, the $\eta$ parameter in Equation 3.5 was chosen to 0.45.

In this paper, we used the ensemble open loop runs, and the deterministic run as benchmarks to evaluate DA performance. In open loop runs, the top layer soil moisture and forcing data were perturbed, and the mean of the ensemble was used as a solution. The PCR-GLOBWB model was spun up from 1st January 2000 to 31st March 2015. The assimilation began from 31st March 2015 to 31st December 2016 with respect to available SMAP data and validation data. All experiments were conducted on the DAS-5 supercomputer [43].

### 5.3.3. Evaluation metrics

Different metrics were used to evaluate and compare the performance of various data assimilation experiments. The accuracy of daily discharge estimates is evaluated using the Nash-Sutcliffe efficiency (NSE) [44], which was calculated following Eq.(5.1):

$$NSE = 1 - \frac{\sum_{k=1}^{N}\left(Q_{\text{obs}}(k) - Q_{\text{sim}}(k)\right)^2}{\sum_{k=1}^{N}\left(Q_{\text{obs}}(k) - \overline{Q_{\text{obs}}}\right)^2} \tag{5.1}$$

Where $Q_{\text{obs}}, Q_{\text{sim}}, \overline{Q_{\text{obs}}}$ are the observed discharge, simulated discharge, and the mean of the observed discharge, respectively. $N$ is the length of the time series and $k$ indicates each time step. $NSE$ ranges from minus infinity to 1, and when the value of $NSE$ is close to 1, it indicates a good match of the estimated discharge to the observed discharge.

The normalized root mean squared error (NRMSE) is also used to quantify the improvement. It was calculated using Eq.(5.2).

$$NRMSE = \frac{\sqrt{\frac{1}{N}\sum_{k=1}^{N}\left(Q_{\text{DA}}(k) - Q_{\text{obs}}(k)\right)^2}}{Q_{\text{obs}}^{\max} - Q_{\text{obs}}^{\min}} \tag{5.2}$$

where $Q_{\text{DA}}$ is discharge estimated after by data assimilation. $Q_{\text{obs}}^{\max}$ and $Q_{\text{obs}}^{\min}$ are the maximum and minimum values of observed discharge, respectively. Additionally, the Pearson correlation coefficient (r) is used to measure the linear relationship between the simulations and observations. All evaluation metrics were assessed for the deterministic simulation, the open loop simulation, and the mean of particles provided by data assimilation.

## 5.4. Results and discussion

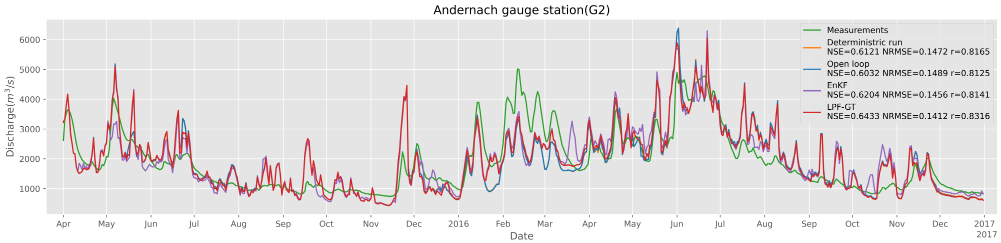

Figure 5.2: Simulated and observed discharge estimates at Andernach gauge station (G2) for the period April 2015-December 2016. It shows a comparison of the time series of discharge estimates obtained from measurements and four simulated experiments, including the deterministic run, the open loop run, and two DA runs: LPF-GT and EnKF. In the legend, $NSE$, $NRMSE$, and $r$ of these experiments were shown separately. Both DA runs used five particles (or ensemble members in EnKF). For LPF-GT settings, the localization scale was 0.12.

Figure 5.2 displayed the discharge time series at the Andernach gauge station(G2) for the deterministic run, one open loop run, and two DA experiments using LPF-GT and EnKF, respectively. In both the open loop and the DA runs, only five particles(or five ensemble members) were applied. Although the hydrological model used in the present study did not need a calibration, the deterministic run still had a high performance, leaving little room for any data assimilation to improve the discharge estimates.

According to the values of $NRMSE$, $NSE$, and $r$ from the open loop run, it indicated that the open run reduced estimation accuracy. It is plausibly because of the sample errors introduced by only using a few particles. According to results obtained from two DA experiments shown in Figure 5.2, the $NSE$ increments and $NRMSE$ reductions from DA runs were higher with respect to the deterministic run. The correlation values $r$ was also increased. Thus, both DA algorithms improved discharge accuracy only using five particles. The $NSE$, $RMSE$, and $r$ from LPF-GT were 0.6433, 0.1412 and 0.8316, respectively. Both $NSE$ and $r$ were higher than EnKF's, and $NRMSE$ had a lower value in comparison, indicating that LPF-GT outperformed EnKF. All values of $NSE$, $NRMSE$, and $r$ from deterministic and open loop runs for all validation stations were provided in Appendix A.

Next, to further investigate the impact of the number of particles on DA performance, we conducted DA experiments with 5, 10, 15, and 20 particles. As the evaluation metrics' judgment is not straightforward to show the improvement or degeneration, we calculated the percentage difference of $NRMSE$,$NSE$ and $r$ in the following experiments. Given space constraints, we only presented results from G2, G5 and G8 stations, and the remaining results were shown in the Appendix A.

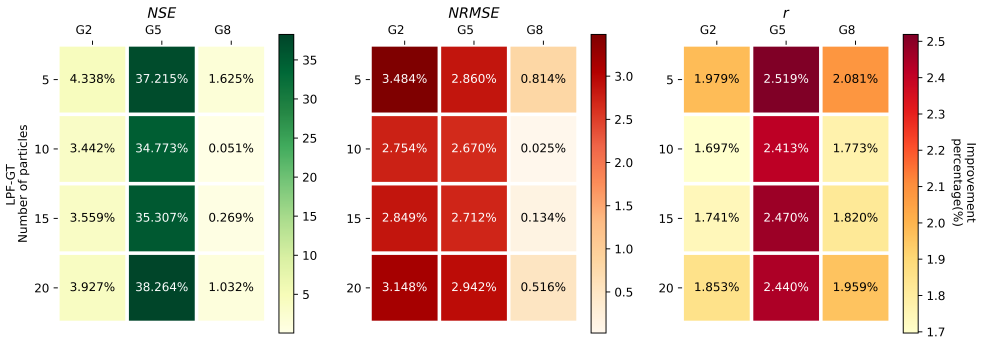

Figure 5.3: Improvement percentage in terms of Nash-Sutcliffe index ($NSE$) in the left column, normalized root mean square error ($NRMSE$) in the middle, and Pearson correlation coefficient($r$) in the right column from DA experiments with different numbers of particles or ensemble sizes using LPF-GT(upper panel) and EnKF(lower panel) with respect to the deterministic run at the G2, G5, and G8 gauge stations.

We set the number of particles to 5, 10, 15, and 20 in both LPF-GT and EnKF DA cases to examine the impact of sample errors, and results were shown in Figure 5.3. The meaning of the ensemble size in EnKF is equivalent to the number of particles in LPF-GT. Thus, we also used the number of particles in EnKF for brevity. Compared with results in the deterministic run, all $NSE$ values of DA runs increased, and the improved percentage ranged from 0.05% to 38%. In all cases, $NRMSE$ values decreased by 0.01% to 3%, except for one case using 10 particles at the G8 site. It was obvious that data assimilation had a positive role in improving discharge estimates.

There was no strong tendency among results of DA plotted in Figure 5.3 when using different numbers of particles. Theoretically, using more particles in DA should obtain better performance. But in Figure 5.3, the evaluation scores did not change with the increase of the number of particles. For example, at the G5 station, for LPF-GT, the NSE and NRMSE improvement went up when increasing the number of particles from 10 to 20. But the case with five particles achieved a better performance than using 10 or 15 particles. In terms of $r$, using five particles had the most significant improvement. We believe that there are two possible reasons for this phenomenon. Using a small number of particles increases sample errors. The reduction of sample errors by increasing the number of particles to 20 may not be not significant. Besides, the use of localization in LPF-GT possibly caused imbalance issues. Each model state variable was updated independently in LPF-GT, which broke the consistency in all the model states. In EnKF, the ensemble of all model states was updated as a whole. Thus, imbalance problems in LPF-GT did not exist in EnKF. But we still found that, at the G5 station, the case with 15 particles gave a larger improvement in $NSE$ and $NRMSE$ than using 20 particles. Similarly, the $NSE$ improvement of five particles was larger than ten particles at the G2 station.

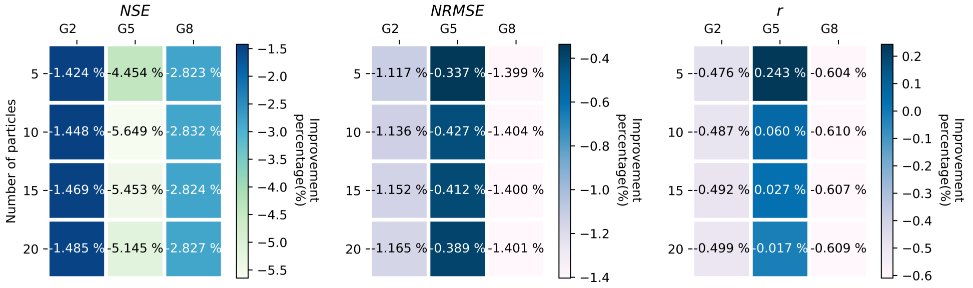

Figure 5.4: Improvement percentage in terms of Nash-Sutcliffe index ($NSE$) and normalized root mean square error ($NRMSE$) from the open loop run without data assimilation with respect to results of the deterministic run at the G2, G5, and G8 gauge stations.

Figure 5.4 shows the increase or decline percentages of $NSE$ and $NRMSE$ for the open loop cases with various settings of the number of particles. The negative values in Figure 5.4 referred to degeneration instead of improvement. It was clear that almost all values of $NSE$ and $NRMSE$ from open loop runs were smaller than values in the deterministic experiment, as they decreased from 0.017% to around 5%. It suggested that the open loop runs with different numbers of particles performed more poorly than the results obtained from the deterministic run. With the growth of the number of particles, the results of the open loop run should have an improvement. Unfortunately, we did not observe the expected results. Considering we only increased particles from 5 to 20, the impact of sample errors brought by a few particles is highly likely to be large. In this case, we believe sample errors are the dominant source of uncertainty, leading to the deterioration of open loop runs consequently.

DA gave a considerably larger improvement than performance in the open loop, which was not shown directly. In the following experiments, we only collected results from DA and deterministic runs for comparisons. In our view, it was unnecessary to compare DA results with open loop runs because of the poor quality of the open loop runs.

A sensitivity analysis was carried out to determine the impact of localization scales on data assimilation performance. As we discussed before, sample errors had an

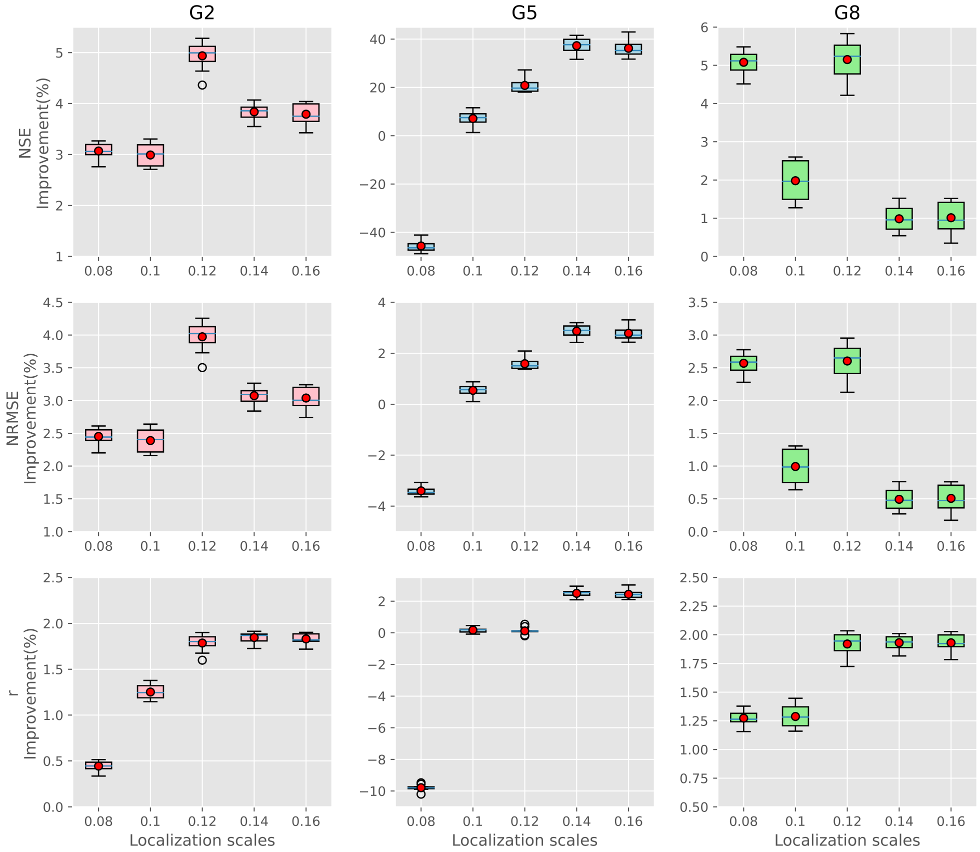

Figure 5.5: Boxplots of improvement percentage in terms of Nash-Sutcliffe index ($NSE$) in the upper row, normalized root mean square error ($NRMSE$) in the middle row, and Pearson correlation coefficient ($r$) in the lower row from LPF-GT and EnKF experiments at the G2 (first column), G5 (middle column), and G8 (right column) gauge stations. In the cases with LPF-GT, five localization scales ranged from 0.08 to 0.16 were used. Results given by EnKF were demonstrated separately at the right side of each LPF-GT run. Each boxplot's results were from one experiment with ten repeating times, and the red point in each boxplot indicates the mean of ten repeating experiments.

impact on LPF-GT, which was not negligible. To avoid the influence of random errors, we repeated each case for LPF-GT with a specific localization scale ten times. Results produced by data assimilation experiments from the three gauge stations are shown in Figure 5.5.

At the G2 station, according to the boxplots of $NSE$, $r$, and $NRMSE$, the impact of sample errors on DA estimates was the smallest, in comparison to results at the other two stations. There was a significant improvement when the localization scale was increased to 0.12. Increasing the value of the localization scale means assimilating more observations in DA. It seemed that when localization scales were set to larger than 0.1, the change of $NSE$ was small, which indicated that the effect of localization is minimal. The same pattern could be observed in $r$ and $NRMSE$. In this case, assimilation with a bigger localization radius leads to a more significant improvement.

Unlike the case in the G2 station, the improvement percentage at the G5 station had a different change pattern with the growth of localization scales. The improvement for $NSE$, $r$, and $NRMSE$ reached a plateau when the localization scale was larger than 0.12. Thus, the use of a relatively large localization scale produced better estimates at the G2 station. But in the case with the localization scale of 0.08, it was apparent that discharge estimates did not benefit from LPF-GT, which probably was caused by insufficient observations within the localization scale. Increasing the localization scale at the G8 station had a negative impact on DA performance. Using the localization scale equal to or bigger than 0.14, there was no noticeable improvement in either $NSE$ or $NRMSE$. But the $r$ value improved slightly. Unlike the situations at in the last two gauge stations, smaller localization scales brought more benefits.

It should be noted that including more observations via increasing localization radius in a DA method does not always bring an improvement. Tuning localization scales to find an optimal value is crucial and inevitable in this method.

In the next assessment, to compare the performance of LPF-GT with EnKF, estimates from both methods of discharge were compared to in situ measurements from gauge stations. In addition, we investigated the impact of observation uncertainty on both DA algorithms' performance. The default standard deviation of observations was 0.04, and all DA experiments were also carried out when setting the standard deviation of observations to 0.05 and 0.06. All related results are shown in Figure 5.6. To avoid the impact of sample errors, all tests were repeated ten times. The results with the best evaluation scores were chosen. Only the $NRMSE$ improvement was given, and the results related to $NSE$ and $r$ were provided in the Appendix A.

Figure 5.6 showed that LPF-GT had a bigger improvement and was more stable than EnKF when observations with more uncertainty were assimilated, except for the cases at the G3 station. It was clear that the fluctuation of LPF-GT's $NRMSE$ improvement was small. For EnKF, its performance was unstable, and it did not bring any improvement at the G4 and G5 stations based on three evaluation metrics. In general, using observations with a larger standard deviation deteriorated EnKF's estimates. But at the G5 station, in the case with the standard deviation of 0.06, LPF-GT gave the most significant improvement. It was likely that the assumption of observation uncertainty at these places was not correct. The true value of the observations' standard deviation was closer to 0.06. Only at the G3 station, EnKF had a better performance than LPF-GT. For the others, LPF-GT outperformed EnKF.

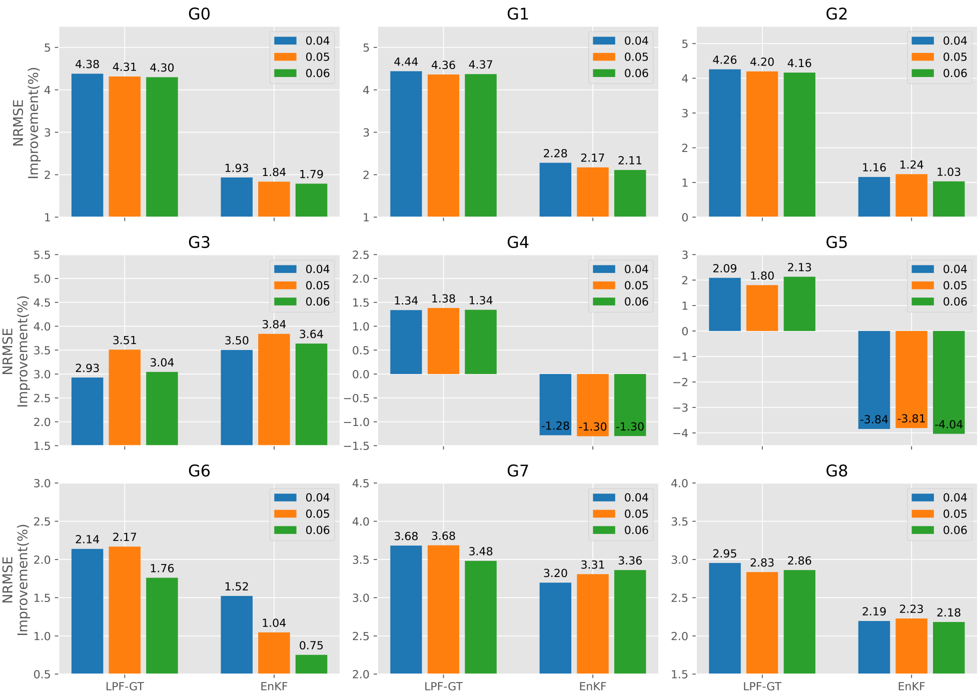

Figure 5.6: Improvement percentage in terms of normalized root mean square error ($NRMSE$) from DA experiments with different observations' standard deviations using LPF-GT and EnKF with respect to the deterministic run at all validation gauge stations.

Theoretically, increasing the uncertainty of observations is equivalent to reducing observations' accuracy, which can decrease DA estimates' accuracy. The impact of observations' uncertainty on EnKF was negative. But, LPF-GT performed more stably with reasonable accuracy, and it was more capable of obtaining useful information from observations. The true uncertainty in observations is unknown, and the assumption about it is often flawed. According to the results in Figure 5.6, further investigation into observation uncertainty is needed.

## 5.5. Conclusions

This study aimed to test if the LPF-GT data assimilation method is usefull for hydrology by assimilating SMAP soil moisture observations into the large-scale PCR-GLOWB hydrological model and compare discharge time series estimates in the Rhine river basin. Particle filters are under-explored as a data assimilation method in hydrological modelling. To show the benefits of LPF-GT, we used EnKF as a benchmark to compare its performance with LPF-GT's. In situ discharge observations from stream gauge stations were used to evaluate improvements brought by data assimilation. This study confirmed that simulated discharge estimates could be improved with the soil moisture assimilation, which had been proven by previous studies. The successful application of soil moisture assimilation scheme via LPF-GT in a high dimensional model proved that this DA algorithm was stable. This research demonstrates that LPF-GT avoided the curse of dimensionality in a real application.

Using five particles in LPF-GT to obtain a satisfactory performance is one benefit we demonstrated in this paper. The low number of particles used in DA can save a lot of computing time. Due to localization applied in LPF-GT, the update for each model state is independent. Thus, LPF-GT can be implemented in parallel, which is one of the advantages of using localization. For models with a larger scale, ie. with more model states, LPF-GT has the ability to improve computing efficiency further. In general, a DA algorithm using more particles or more ensemble members can yield estimates with higher accuracy. On the contrary, few particles bring more sample errors to the DA system. Results from the open loop run show poor PCR-GLOBWB discharge estimates, probably caused by sample errors. Besides, sample errors could make DA unstable based on the results shown in Figure 5.5. Reducing sample errors definitely leads to further improvement of DA.

Tuning localization scales to find an optimal value is inevitable. Results indicated that the optimal localization radius is not a global constant. It varies in different cases. The tuning process takes time, which is the disadvantage brought by localization. In addition, localization has imbalance issues, probably resulting in limiting DA improvement. Developing adaptive localization methods is a possible way to solve the imbalance issues. We also found that LPF-GT performed stably when changing the observations' standard deviation. LPF-GT has the ability to use more useful information from observations. The real uncertainty of observations is generally unknown. In this case, LPF-GT is a better choice with stable performance.

In conclusion, the successful assimilation of SMAP soil moisture retrievals to improve model estimates of discharge over a sizeable spatial domain via LPF-GT verified the possibility of applying particle filters in hydrological data assimilation at large scales. LPF-GT can be used in other types of basins to examine how discharge estimates can benefit from particle filters with localization. LPF-GT should also be adapeted to assimilate other observations into hydrological models to improve other model components.

## References

[1] P. López López, N. Wanders, J. Schellekens, L. J. Renzullo, E. H. Sutanudjaja, and M. F. P. Bierkens, *Improved large-scale hydrological modelling through the assimilation of streamflow and downscaled satellite soil moisture observations,* Hydrology and Earth System Sciences 20, 3059 (2016).

[2] J. Loizu, C. Massari, J. Álvarez-Mozos, A. Tarpanelli, L. Brocca, and J. Casalí, *On the assimilation set-up of ASCAT soil moisture data for improving streamflow catchment simulation,* Advances in Water Resources 111, 86 (2018).

[3] S. Azimi, A. B. Dariane, S. Modanesi, B. Bauer-Marschallinger, R. Bindlish, W. Wagner, and C. Massari, *Assimilation of Sentinel 1 and SMAP – based satellite soil moisture retrievals into SWAT hydrological model: The impact of satellite revisit time and product spatial resolution on flood simulations in small basins,* Journal of Hydrology 581, 124367 (2020).

[4] S. Wongchuig, A. Fleischmann, R. Paiva, and A. Fadel, *Toward Discharge Estimation for Water Resources Management with a Semidistributed Model and Local Ensemble Kalman Filter Data Assimilation,* Journal of Hydrologic Engineering 26, 05020047 (2021).

[5] N. J. Gordon, D. J. Salmond, and A. F. M. Smith, *Novel approach to nonlinear/non-Gaussian Bayesian state estimation,* IEE Proceedings F (Radar and Signal Processing) 140, 107 (1993).

[6] A. Doucet, N. de Freitas, and N. Gordon, *An Introduction to Sequential Monte Carlo Methods,* in *Sequential Monte Carlo Methods in Practice*, Statistics for Engineering and Information Science (Springer, New York, NY, 2001) pp. 3–14.

[7] Z. Wang, R. Hut, and N. Van de Giesen, *A Local Particle Filter Using Gamma Test Theory for High-Dimensional State Spaces,* Journal of Advances in Modeling Earth Systems 12, e2020MS002130 (2020).

[8] S. I. Seneviratne, T. Corti, E. L. Davin, M. Hirschi, E. B. Jaeger, I. Lehner, B. Orlowsky, and A. J. Teuling, *Investigating soil moisture–climate interactions in a changing climate: A review,* Earth-Science Reviews 99, 125 (2010).

[9] S. V. Kumar, C. D. Peters-Lidard, J. A. Santanello, R. H. Reichle, C. S. Draper, R. D. Koster, G. Nearing, and M. F. Jasinski, *Evaluating the utility of satellite soil moisture retrievals over irrigated areas and the ability of land data assimilation methods to correct for unmodeled processes,* Hydrology and Earth System Sciences 19, 4463 (2015).

[10] C. Massari, L. Brocca, A. Tarpanelli, and T. Moramarco, *Data Assimilation of Satellite Soil Moisture into Rainfall-Runoff Modelling: A Complex Recipe?* Remote Sensing 7, 11403 (2015).

[11] E. M. Demaria, B. Nijssen, and T. Wagener, *Monte Carlo sensitivity analysis of land surface parameters using the Variable Infiltration Capacity model,* Journal of Geophysical Research: Atmospheres 112 (2007), 10.1029/2006JD007534.

[12] Y. H. Kerr, P. Waldteufel, P. Richaume, J. P. Wigneron, P. Ferrazzoli, A. Mahmoodi, A. A. Bitar, F. Cabot, C. Gruhier, S. E. Juglea, D. Leroux, A. Mialon, and S. Delwart, *The SMOS Soil Moisture Retrieval Algorithm,* IEEE Transactions on Geoscience and Remote Sensing 50, 1384 (2012).

[13] D. Entekhabi, E. G. Njoku, P. E. O'Neill, K. H. Kellogg, W. T. Crow, W. N. Edelstein, J. K. Entin, S. D. Goodman, T. J. Jackson, J. Johnson, J. Kimball, J. R. Piepmeier, R. D. Koster, N. Martin, K. C. McDonald, M. Moghaddam, S. Moran, R. Reichle, J. C. Shi, M. W. Spencer, S. W. Thurman, L. Tsang, and J. V. Zyl, *The Soil Moisture Active Passive (SMAP) Mission,* Proceedings of the IEEE 98, 704 (2010).

[14] M. Owe, R. de Jeu, and T. Holmes, *Multisensor historical climatology of satellite-derived global land surface moisture,* Journal of Geophysical Research: Earth Surface 113 (2008), 10.1029/2007JF000769.

[15] V. Naeimi, K. Scipal, Z. Bartalis, S. Hasenauer, and W. Wagner, *An Improved Soil Moisture Retrieval Algorithm for ERS and METOP Scatterometer Observations,* IEEE Transactions on Geoscience and Remote Sensing 47, 1999 (2009).

[16] N. Tangdamrongsub, S. C. Steele-Dunne, B. C. Gunter, P. G. Ditmar, and A. H. Weerts, *Data assimilation of GRACE terrestrial water storage estimates into a regional hydrological model of the Rhine River basin,* Hydrology and Earth System Sciences 19, 2079 (2015).

[17] H. Lievens, S. K. Tomer, A. Al Bitar, G. J. M. De Lannoy, M. Drusch, G. Dumedah, H. J. Hendricks Franssen, Y. H. Kerr, B. Martens, M. Pan, J. K. Roundy, H. Vereecken, J. P. Walker, E. F. Wood, N. E. C. Verhoest, and V. R. N. Pauwels, *SMOS soil moisture assimilation for improved hydrologic simulation in the Murray Darling Basin, Australia,* Remote Sensing of Environment 168, 146 (2015).

[18] N. Tangdamrongsub, S. C. Steele-Dunne, B. C. Gunter, P. G. Ditmar, E. H. Sutanudjaja, Y. Sun, T. Xia, and Z. Wang, *Improving estimates of water resources in a semi-arid region by assimilating GRACE data into the PCR-GLOBWB hydrological model,* Hydrology and Earth System Sciences 21, 2053 (2017).

[19] N. Wanders, M. F. P. Bierkens, S. M. de Jong, A. de Roo, and D. Karssenberg, *The benefits of using remotely sensed soil moisture in parameter identification of large-scale hydrological models,* Water Resources Research 50, 6874 (2014).

[20] N. Tangdamrongsub, S.-C. Han, I.-Y. Yeo, J. Dong, S. C. Steele-Dunne, G. Willgoose, and J. P. Walker, *Multivariate data assimilation of GRACE, SMOS, SMAP measurements for improved regional soil moisture and groundwater storage estimates,* Advances in Water Resources 135, 103477 (2020).

[21] S. Meng, X. Xie, and S. Liang, *Assimilation of soil moisture and streamflow observations to improve flood forecasting with considering runoff routing lags,* Journal of Hydrology 550, 568 (2017).

[22] C. M. Emery, A. Paris, S. Biancamaria, A. Boone, S. Calmant, P.-A. Garambois, and J. S. da Silva, *Large-scale hydrological model river storage and discharge correction using a satellite altimetry-based discharge product,* Hydrology and Earth System Sciences 22, 2135 (2018).

[23] C. Massari, S. Camici, L. Ciabatta, L. Brocca, C. Massari, S. Camici, L. Ciabatta, and L. Brocca, *Exploiting Satellite-Based Surface Soil Moisture for Flood Forecasting in the Mediterranean Area: State Update Versus Rainfall Correction,* Remote Sensing 10, 292 (2018).

[24] G. Evensen, *The Ensemble Kalman Filter: Theoretical formulation and practical implementation,* Ocean Dynamics 53, 343 (2003).

[25] P. L. Houtekamer and F. Zhang, *Review of the Ensemble Kalman Filter for Atmospheric Data Assimilation,* Monthly Weather Review 144, 4489 (2016).

[26] I. Hoteit, X. Luo, and D.-T. Pham, *Particle Kalman Filtering: A Nonlinear Bayesian Framework for Ensemble Kalman Filters,* Monthly Weather Review 140, 528 (2011).

[27] T. Bengtsson, P. Bickel, and B. Li, *Curse-of-Dimensionality Revisited: Collapse of the Particle Filter in Very Large Scale Systems* (Institute of Mathematical Statistics, 2008).

[28] C. Snyder, T. Bengtsson, P. Bickel, and J. Anderson, *Obstacles to High-Dimensional Particle Filtering,* Monthly Weather Review 136, 4629 (2008).

[29] P. Rebeschini and R. van Handel, *Can local particle filters beat the curse of dimensionality?* The Annals of Applied Probability 25, 2809 (2015).

[30] S. Robert and H. R. Künsch, *Localizing the Ensemble Kalman Particle Filter,* Tellus A: Dynamic Meteorology and Oceanography 69, 1282016 (2017).

[31] A. Farchi and M. Bocquet, *Review article: Comparison of local particle filters and new implementations,* Nonlinear Processes in Geophysics 25, 765 (2018).

[32] J. Poterjoy, *A Localized Particle Filter for High-Dimensional Nonlinear Systems,* Monthly Weather Review 144, 59 (2016).

[33] S. G. Penny and T. Miyoshi, *A local particle filter for high dimensional geophysical systems,* Nonlinear Processes in Geophysics Discussions 2, 1631 (2015).

[34] E. H. Sutanudjaja, R. van Beek, N. Wanders, Y. Wada, J. H. C. Bosmans, N. Drost, R. J. van der Ent, I. E. M. de Graaf, J. M. Hoch, K. de Jong, D. Karssenberg, P. López López, S. Peßenteiner, O. Schmitz, M. W. Straatsma, E. Vannametee, D. Wisser, and M. F. P. Bierkens, *PCR-GLOBWB 2: A 5&thinsp;arcmin global hydrological and water resources model,* Geoscientific Model Development 11, 2429 (2018).

[35] S. Khanal, A. F. Lutz, W. W. Immerzeel, H. de Vries, N. Wanders, and B. van den Hurk, *The Impact of Meteorological and Hydrological Memory on Compound Peak Flows in the Rhine River Basin,* Atmosphere 10, 171 (2019).

[36] D. Viviroli, R. Weingartner, and B. Messerli, *Assessing the Hydrological Significance of the World's Mountains,* Mountain Research and Development 23, 32 (2003).

[37] C. S. Photiadou, A. H. Weerts, and B. J. J. M. van den Hurk, *Evaluation of two precipitation data sets for the Rhine River using streamflow simulations,* Hydrology and Earth System Sciences 15, 3355 (2011).

[38] M. R. Haylock, N. Hofstra, A. M. G. K. Tank, E. J. Klok, P. D. Jones, and M. New, *A European daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006,* Journal of Geophysical Research: Atmospheres 113 (2008), 10.1029/2008JD010201.

[39] W. Hamon, *Computation of direct runoff amounts from storm rainfall,* International Association of Scientific Hydrology Publication 63, 52 (1963).

[40] Y. Wada, L. P. H. van Beek, D. Viviroli, H. H. Dürr, R. Weingartner, and M. F. P. Bierkens, *Global monthly water stress: 2. Water demand and severity of water stress,* Water Resources Research 47 (2011), 10.1029/2010WR009792.

[41] P. Liu, J. Judge, R. D. D. Roo, A. W. England, and T. Bongiovanni, *Uncertainty in Soil Moisture Retrievals Using the SMAP Combined Active–Passive Algorithm for Growing Sweet Corn,* IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 9, 3326 (2016).

[42] A. Colliander, T. J. Jackson, R. Bindlish, S. Chan, N. Das, S. B. Kim, M. H. Cosh, R. S. Dunbar, L. Dang, L. Pashaian, J. Asanuma, K. Aida, A. Berg, T. Rowlandson, D. Bosch, T. Caldwell, K. Caylor, D. Goodrich, H. al Jassar, E. Lopez-Baeza, J. Martínez-Fernández, A. González-Zamora, S. Livingston, H. McNairn, A. Pacheco, M. Moghaddam, C. Montzka, C. Notarnicola, G. Niedrist, T. Pellarin, J. Prueger, J. Pulliainen, K. Rautiainen, J. Ramos, M. Seyfried, P. Starks, Z. Su, Y. Zeng, R. van der Velde, M. Thibeault, W. Dorigo, M. Vreugdenhil, J. P. Walker, X. Wu, A. Monerris, P. E. O'Neill, D. Entekhabi, E. G. Njoku, and S. Yueh, *Validation of SMAP surface soil moisture products with core validation sites,* Remote Sensing of Environment 191, 215 (2017).

[43] H. Bal, D. Epema, C. de Laat, R. van Nieuwpoort, J. Romein, F. Seinstra, C. Snoek, and H. Wijshoff, *A Medium-Scale Distributed System for Computer Science Research: Infrastructure for the Long Term,* Computer 49, 54 (2016).

[44] J. E. Nash and J. V. Sutcliffe, *River flow forecasting through conceptual models part I — A discussion of principles,* Journal of Hydrology 10, 282 (1970).

# 6

# Conclusions

*"We are all in the gutter, but some of us are looking at the stars."*

Oscar Wilde, Lady Windermere's Fan

*This dissertation has introduced new additions to particle filters in data assimilation. These additions intend to improve our accurate estimation of uncertainty and to come up with new solutions to problems introduced by nonlinearity in the DA system. To overcome the curse of dimensionality of particle filters, two localization methods were examined separately. To account for uncertainty brought by data assimilation, the Gamma test theory was applied (LPF-GT). To cope with nonlinearity introduced by nonlinear observation operators, Gaussian process regression (LPF-GPR) was used. To explore the possibility of LPF-GT, it has been evaluated in a real hydrological application to improve the accuracy of discharge estimates. With all results presented and discussed in the previous chapters, this dissertation's primary objective mentioned in Chapter 1 has been accomplished. The main conclusions and insights as well as opportunities for further research are summarized in this chapter.*

## 6.1. Main original contributions

### 6.1.1. On introduced data assimilation algorithms

Two local particle filters have been proposed and evaluated using the classical toy model - Lorenz96 in this dissertation, which are the main scientific contributions of this thesis. Both local particle filters achieve a better performance than the benchmark LETKF when using nonlinear observation operators in a series of experiments. It is not surprising to find that particle filters outperform Ensemble-type filters under nonlinear conditions. The main reason for it is that the nonlinearity introduced by observation operators breaks the linear and Gaussian assumptions, which are the theoretical foundation of LETKF. Consequently, LETKF yields sub-optimal and poor estimations. Particle filters are more capable of extracting information from nonlinear cases mainly because they do not rely on linear and Gaussian assumptions.

#### On LPF-GT algorithm

The Local Particle Filter with Gamma Test Theory (LPF-GT) is introduced in chapter 3, which is based on Wang *et al.* [1]. In LPF-GT, two related strategies are applied to avoid filter collapse and to achieve acceptable results.

1. State-domain localization is used. The update for each model state relies on an independent analysis, and only nearby observations within the localization scale are assimilated.

2. The addition of the Gamma Test Theory (LPF-GT) addresses the problem of underestimation of the uncertainty of the state by explicitly considering the added uncertainty of the data assimilation step, thus preventing filter collapse when using a low number of particles.

   As discussed in Chapter 3, the Gamma test theory is a viable tool to identify the potential source of uncertainty brought by data assimilation, thereby improving data assimilation performance. It should be noted that data assimilation updates state variables based on uncertainty information of observations, but it can bring extra uncertainty into the system. Besides, quantifying uncertainty in nonlinear observations enables us to understand observations better and allows us to extract more useful information from them.

#### On LPF-GPR algorithm

Similar to LPF-GT, LPF-GPR introduces two more approaches to overcome the curse of dimensionality and to obtain satisfactory performance.

1. Localization is also applied in LPF-GPR. Nevertheless, unlike LPF-GT, the specific localization method used is called sequential-observation localization. All observations are assimilated one by one, and one observation only has an influence on the model states within localization radius. The location of each observation is the center point when choosing model states based on localization radius.

2. Replacing the observation operator with surrogate models based on Gaussian Process Regression (LPF-GPR) allows accounting for nonlinearity in the transformation from model state to observations, a frequently encountered issue in

geoscientific observations. As shown in Chapter 4, the alternates of observation operators provide us a way to deal with nonlinearity in data assimilation from a different angle.

### 6.1.2. On satellite soil moisture data assimilation

Soil moisture is a crucial hydrological component to many fields such as agricultural water management, flood prediction, and land-atmosphere modeling. But simulating soil moisture, discharge, and other water fluxes accurately remains a challenge. In previous studies, assimilation of satellite soil moisture to improve discharge estimates has been confirmed via the popular DA algorithm: Ensemble Kalman Filter (EnKF). New proposed local particle filters(LPF-GT) in Chapter 3 have been applied to facilitate the assimilation of SMAP satellite soil moisture products in chapter 5 of this dissertation. DA applications in hydrology using particle filters are rare in comparison to other assimilation studies. The potential of recently developed LPF-GT was explored for the first time in a hydrological application for discharge estimation.

In this thesis, using only five particles in DA runs, the newly introduced methods could achieve satisfactory performance. The success of applying LPF-GT proved the capacity and the stability of this algorithm. A series of sensitivity experiments suggested that some factors could have an adverse impact on LPF-GT's performance. The first was the sampling error introduced by a few particles. The benefits of using a small number of particles are immediately apparent because less compute resources are needed. But the bias caused by five particles was inevitable. The collective impact of sampling error and imbalance resulting from localization is the main reason for the phenomenon, in which the results did not become better with the increase in the number of particles. Besides, there existed an optimal localization scale for a specific case. Tuning this parameter for better results is necessary. All these findings are instructive and informative for further DA applications with satellite data. Lastly, LPF-GT performs stably when increasing the uncertainty of observations.

## 6.2. Future research

### 6.2.1. Adaptive localization methods

Localization has been proven an effective solution to the curse of dimensionality of particle filters in high dimensional systems. However, finding an appropriate localization scale, which provides the best data assimilation estimates, is challenging. In this thesis, proper localization scales were found by trial and error, and we have to admit that it was time-consuming. After setting several localization scales within a specific range in experiments, the best result indicates the suitable localization scale. It should be noted that the most proper localization does not exist because we found that experiments with different settings for localization scales could have a similar performance.

Thus, it is absolutely essential to develop adaptive localization methods. In recent years, for the Ensemble-type filters, several researchers have attempted to find factors influencing the optimal localization radius [2, 3], and some adaptive localization methods have been proposed [4–7]. All these methods have a certain ability to specify the localization radius adaptively, but they still need tuning. Although these academic findings provide some insights for particle filters, studies of adaptive localization methods for particle filters are rare.

### 6.2.2. Potential surrogate models for nonlinear observation operators

The nature of using the Gaussian progress regression models is to change the non-Gaussian uncertainty into Gaussian. We could find massive alternates as surrogate models to replace the nonlinear observation operator in the Machine learning field, like various artificial neural networks. The reason why the Gaussian process regression method is chosen as the surrogate model for the observation operator is that it has a suitable property, and it can give estimates and related uncertainty information at the same time. As long as a model has the same or similar property or the surrogate model's uncertainty is easy to know, this method can become the replacement for nonlinear observation operators. Considering this, Bayesian neural networks are a promising and encouraging candidate. It can take more connected influencing factors into consideration, probably giving comprehensive information about more than one source of uncertainty. Using a surrogate model introduces additional uncertainty inevitably. But the uncertainty information provided by the surrogate model contains both the uncertainty brought by itself and observation errors.

### 6.2.3. Imbalance caused by data assimilation and localization

Data assimilation methods, such as Ensemble Kalman filters and particle filters, are pure mathematical methods. Therefore, it is possible to meet a situation where data assimilation estimates can be beyond the normal range of a model variable. In this case, data assimilation probably violates the dynamical balance of hydrological processes and disturbs water storage and water fluxes in fundamental water balance equations. To address this issue, data assimilation should be constrained by more conditions to keep the consistency between water fluxes and maintain the water balance. Recently, some researchers have started to pay attention to the water balance problem caused by data assimilation. Khaki et al. have proposed new approaches named Weak Constrained Ensemble Kalman Filter (WCEnKF) [8] and unsupervised WCEnKF (UWCEnKF) [9] to cope with the imbalance issue. These methods rely on estimating the covariance associated with the water balance model and other model states. A general framework for this type of imbalance for particle filters and data assimilation needs further attention in the future.

Similarly, localization is another possible source of imbalance in data assimilation because it can break the relationship between state variables. When performing localization at each grid point by removing observations outside the localization radius, all local analyses are independent, which is easy to implement and to parallelize. But this possibly harms the consistency between each model state, and the distribution of all model states is broken. Post-processing model states may mitigate the imbalance issue. Penny and Miyoshi [10] used the deterministic resampling approach of Kitagawa to smooth the transition between nearby grid points. Reducing imbalance brought by localization may further improve estimates given by data assimilation and may eventually lead to more consistent model states.

### 6.2.4. Data assimilation for use in hydrological forecasting with satellite data

Satellite remote sensing data give hydrology new opportunities. They provide diverse types of observations on large scales with high spatial resolutions, revolutionizing how we describe and monitor typical hydrological processes. The ability to assimilate satellite data into a hydrological model to improve some hydrological components has been proved by numerous studies. Data assimilation needs three essential elements: a model propagated by time, related observations, and a data assimilation algorithm. Enhancing the performance of a data assimilation application can be achieved from any or all of these three aspects. For example, data assimilation largely depends on the accurate approximation of uncertainty of the model and observations, and we always make some assumptions about uncertainty. Data assimilation could give better estimates if we have a better understanding of error structure in models and observations. From the aspect of data assimilation algorithms, local particle filters are non-Gaussian filters and require fewer particles. The application prospect of local particle filters and their variants is up-and-coming for other hydrology applications.

## References

[1] Z. Wang, R. Hut, and N. Van de Giesen, *A Local Particle Filter Using Gamma Test Theory for High-Dimensional State Spaces,* Journal of Advances in Modeling Earth Systems 12, e2020MS002130 (2020).

[2] Á. Periáñez, H. Reich, and R. Potthast, *Optimal Localization for Ensemble Kalman Filter Systems,* Journal of the Meteorological Society of Japan. Ser. II 92, 585 (2014).

[3] J. Flowerdew, *Towards a theory of optimal localisation,* Tellus A: Dynamic Meteorology and Oceanography 67, 25257 (2015).

[4] J. L. Anderson, *Exploring the need for localization in ensemble data assimilation using a hierarchical ensemble filter,* Physica D: Nonlinear Phenomena Data Assimilation, 230, 99 (2007).

[5] C. Bishop and D. Hodyss, *Ensemble covariances adaptively localized with ECO-RAP. Part 1: Tests on simple error models,* Tellus A: Dynamic Meteorology and Oceanography 61, 84 (2009).

[6] P. Kirchgessner, L. Nerger, and A. Bunse-Gerstner, *On the Choice of an Optimal Localization Radius in Ensemble Kalman Filter Methods,* Monthly Weather Review 142, 2165 (2014).

[7] N. A. Gasperoni and X. Wang, *Adaptive Localization for the Ensemble-Based Observation Impact Estimate Using Regression Confidence Factors,* Monthly Weather Review 143, 1981 (2015).

[8] M. Khaki, B. Ait-El-Fquih, I. Hoteit, E. Forootan, J. Awange, and M. Kuhn, *A two-update ensemble Kalman filter for land hydrological data assimilation with an uncertain constraint,* Journal of Hydrology 555, 447 (2017).

[9] M. Khaki, B. Ait-El-Fquih, I. Hoteit, E. Forootan, J. Awange,  and M. Kuhn, *Unsupervised ensemble Kalman filtering with an uncertain constraint for land hydrological data assimilation,* Journal of Hydrology 564, 175 (2018).

[10] S. G. Penny and T. Miyoshi, *A local particle filter for high dimensional geophysical systems,* Nonlinear Processes in Geophysics Discussions 2, 1631 (2015).

# Acknowledgements

Above all, I would like to express my most profound appreciation to my daily supervisor, dr. ir. Rolf Hut and promoter Prof. dr. ir. Nick van de Giesen for their expert and patient guidance. Without their help, it is impossible for me to be a scientist.

I would like to give my special thanks to Jianzhi Dong, Yang Lv, and Jianbing Jin, who gave me many support and advice. My understandings of many concepts in data assimilation were obtained by discussions with them. I also want to thank Cong Xiao for helping me get more computing resources, enabling me to finish my experiments using less time. I am grateful to Pro. ir. Susan Steele-Dunne, who explained satellite soil moisture and gave me valuable advice on choosing the right soil moisture products. Besides, I thanks Natthachet Tangdamrongsub for his patient explanation and useful suggestions. Natthachet helped me process some meteorological data, and it saved me a tremendous amount of time.

Many thanks go to my colleagues in the Department of Water Management: Bart Schilperoort, Xin Tian, Jerom Aerts, Yingrong Wen, Wei Shao, Yiqun Song, Yuchen Tang, Dengxiao Lang, Changrang zhou, Xu Shan for all the pleasurable moments we had during my study. I really appreciated assistance from our secretaries Lydia, Betty, and Tamara, in the department.

Heartfelt thanks are given to my Chinese friends, Yan Song, Rui Li, Yan Teng, Kaiyi Zhu in the Netherlands. Because of them, I took countless dinners and other social activities, which made my life in Delft was leisurely and comfortable.

I wish to express my gratitude to the China Scholarship Council(CSC) for providing financial support to give me this great opportunity to continue my research career at the Delft University of Technology.

Finally, I would particularly like to thank my parents for their generous support, unconditional love, and constant encouragement for my Ph.D. study in the Netherlands.

*Zhenwu Wang*
*Delft, March 2021*

# A

# Appendix

*Supplementary material for Chapter 5 is provided in this appendix.*

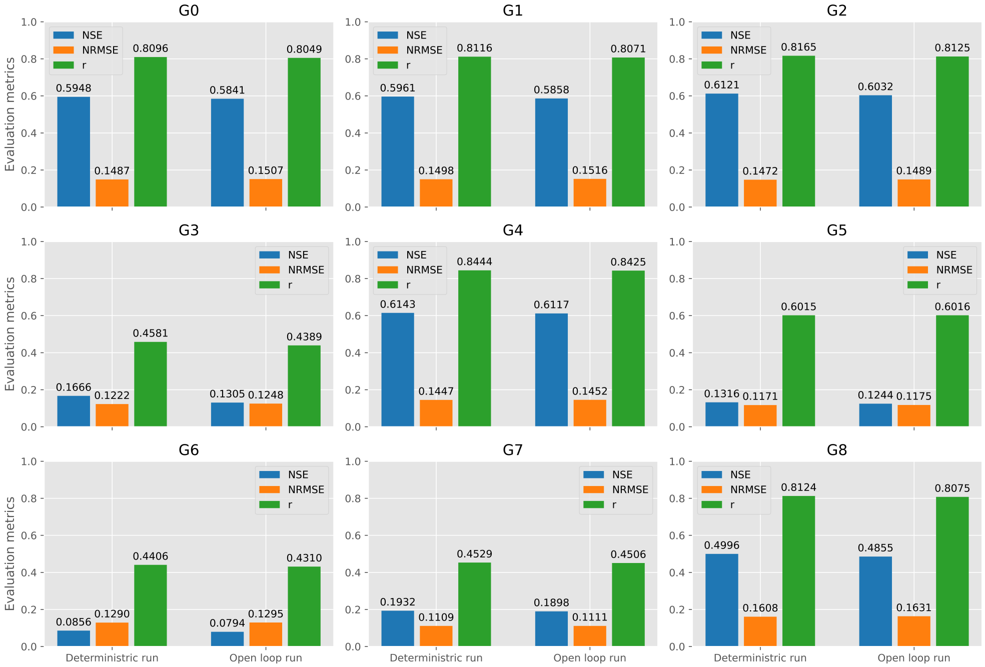

Figure A.1: In terms of Nash-Sutcliffe index ($NSE$), normalized root mean square error ($NRMSE$), and Pearson correlation coefficient ($r$) from deterministic and open loop runs without data assimilation using LPF-GT.

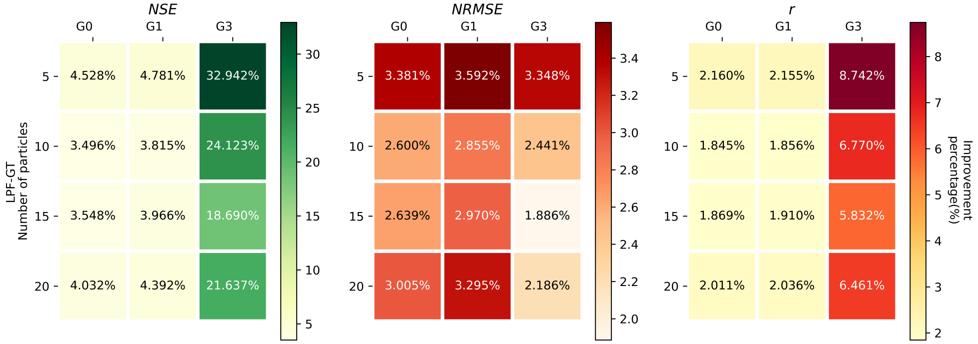

Figure A.2: Improvement percentage in terms of Nash-Sutcliffe index ($NSE$) in the left column, normalized root mean square error ($NRMSE$) in the middle, and Pearson correlation coefficient($r$) in the right column from DA experiments with different numbers of particles or ensemble sizes using LPF-GT (upper panel) and EnKF (lower panel) with respect to the deterministic run at the G0, G1, and G3 gauge stations.

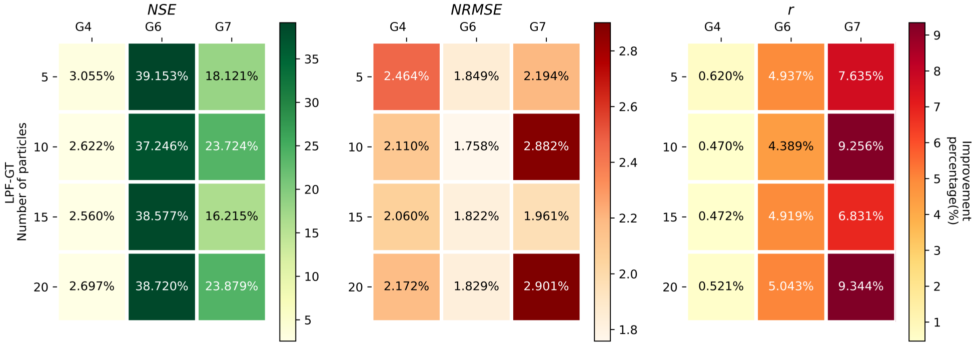

Figure A.3: Improvement percentage in terms of Nash-Sutcliffe index ($NSE$) in the left column, normalized root mean square error ($NRMSE$) in the middle, and Pearson correlation coefficient ($r$) in the right column from DA experiments with different numbers of particles or ensemble sizes using LPF-GT (upper panel) and EnKF (lower panel) with respect to the deterministic run at the G4, G6, and G7 gauge stations.

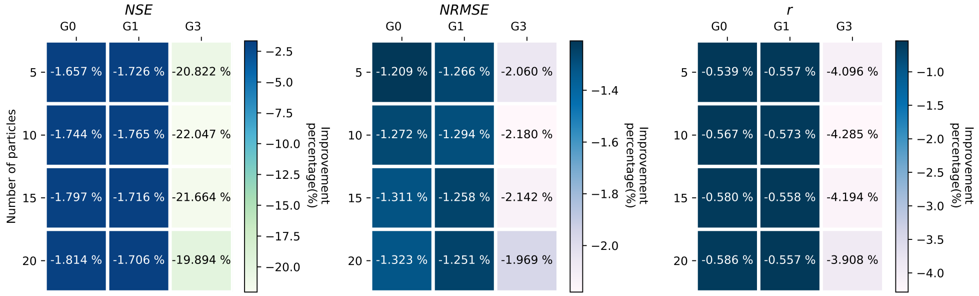

Figure A.4: Improvement percentage in terms of Nash-Sutcliffe index ($NSE$), normalized root mean square error ($NRMSE$) and Pearson correlation coefficient ($r$) from the open loop run without data assimilation with respect to results of the deterministic run at the G0, G1, and G3 gauge stations.

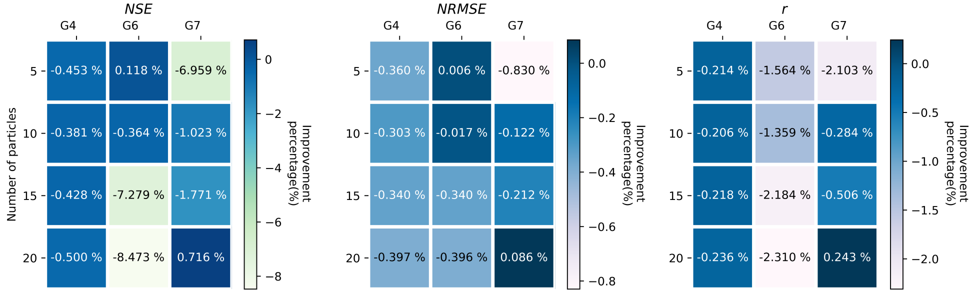

Figure A.5: Improvement percentage in terms of Nash-Sutcliffe index ($NSE$), normalized root mean square error ($NRMSE$) and Pearson correlation coefficient ($r$) from the open loop run without data assimilation with respect to results of the deterministic run at the G4, G6, and G7 gauge stations.

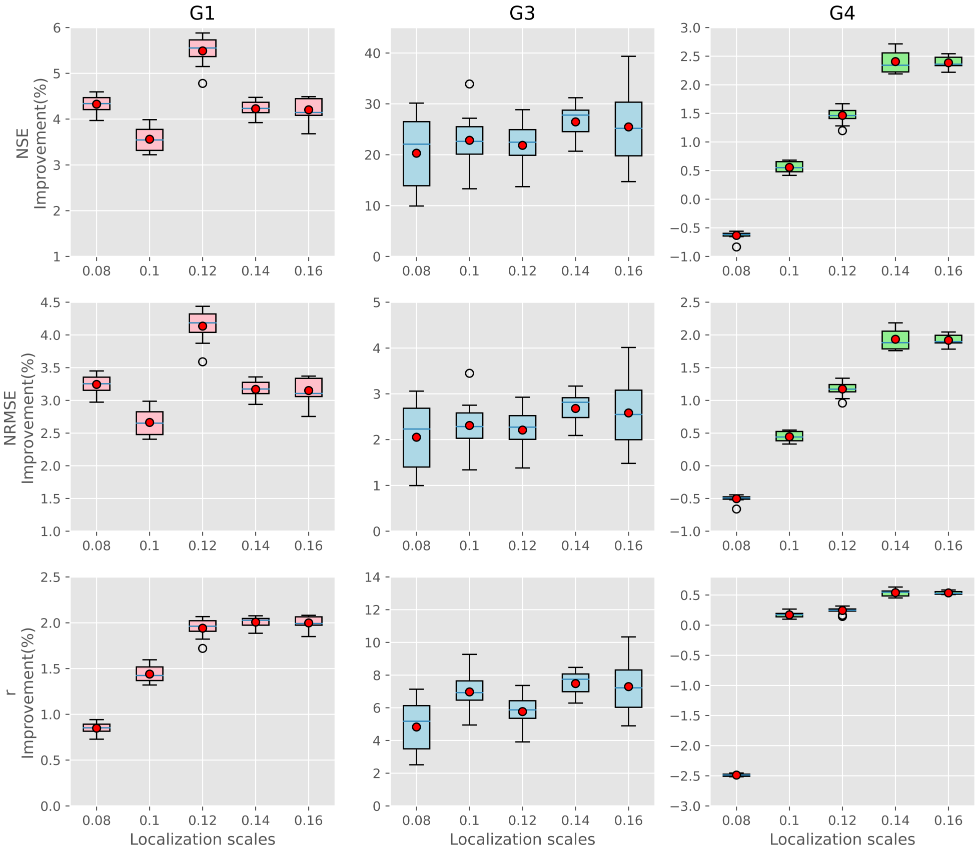

Figure A.6: Boxplots of improvement percentage in terms of Nash-Sutcliffe index ($NSE$) in the upper row, normalized root mean square error ($NRMSE$) in the middle row, and Pearson correlation coefficient ($r$) in the lower row from LPF-GT and EnKF experiments at the G0 (first column), G1 (middle column), and G3 (right column) gauge stations. In the cases with LPF-GT, five localization scales ranged from 0.08 to 0.16 were used. Results given by EnKF were demonstrated separately at the right side of each LPF-GT run. Each boxplot's results were from one experiment with ten repeating times, and the red point in each boxplot indicates the mean of ten repeating experiments.

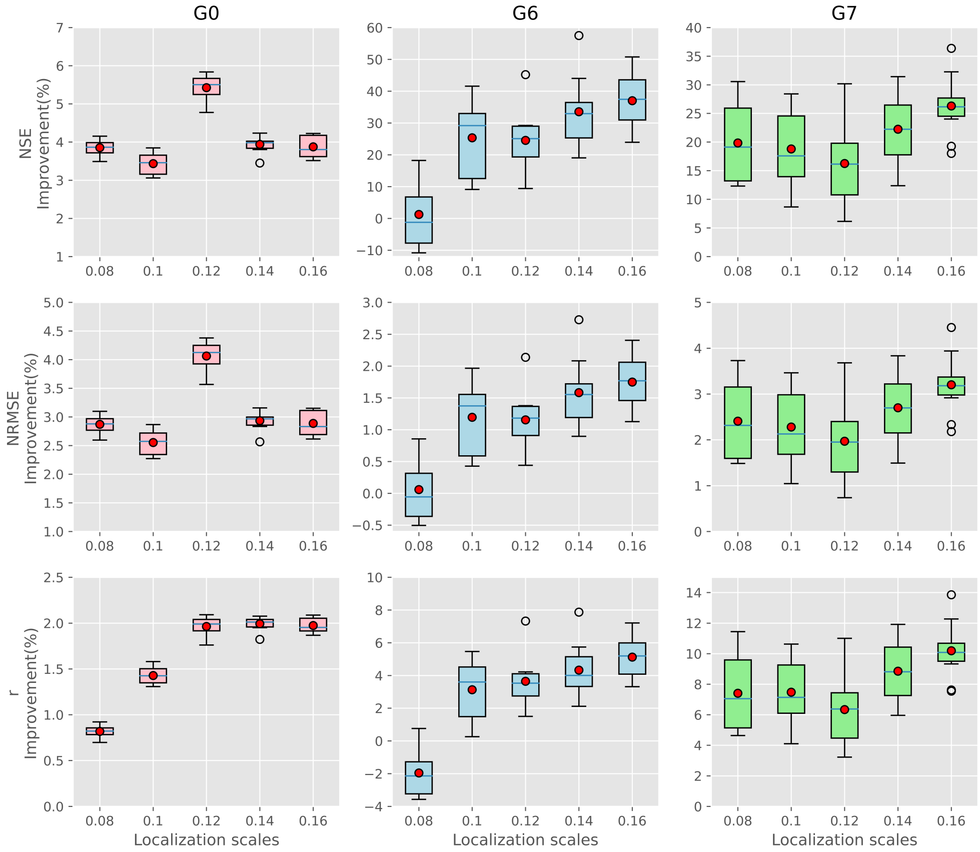

Figure A.7: Boxplots of improvement percentage in terms of Nash-Sutcliffe index ($NSE$) in the upper row, normalized root mean square error ($NRMSE$) in the middle row, and Pearson correlation coefficient ($r$) in the lower row from LPF-GT and EnKF experiments at the G4 (first column), G6(middle column), and G7 (right column) gauge stations. In the cases with LPF-GT, five localization scales ranged from 0.08 to 0.16 were used. Results given by EnKF were demonstrated separately at the right side of each LPF-GT run. Each boxplot's results were from one experiment with ten repeating times, and the red point in each boxplot indicates the mean of ten repeating experiments.

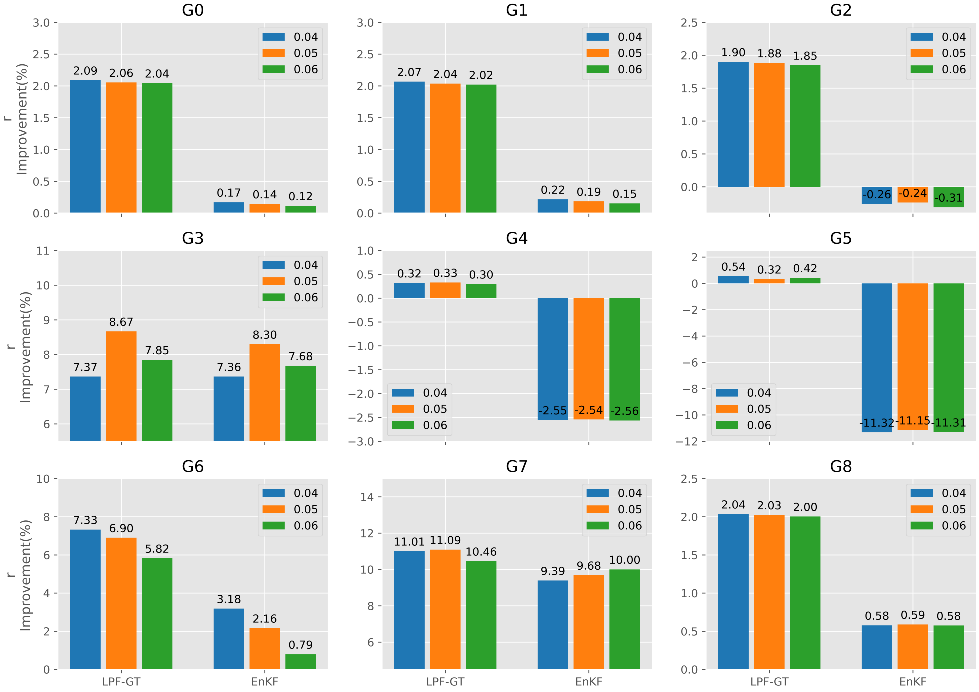

Figure A.8: Improvement percentage in terms of Pearson correlation coefficient ($r$) from DA experiments with different observations' standard deviations using LPF-GT and EnKF with respect to the deterministic run at all validation gauge stations.

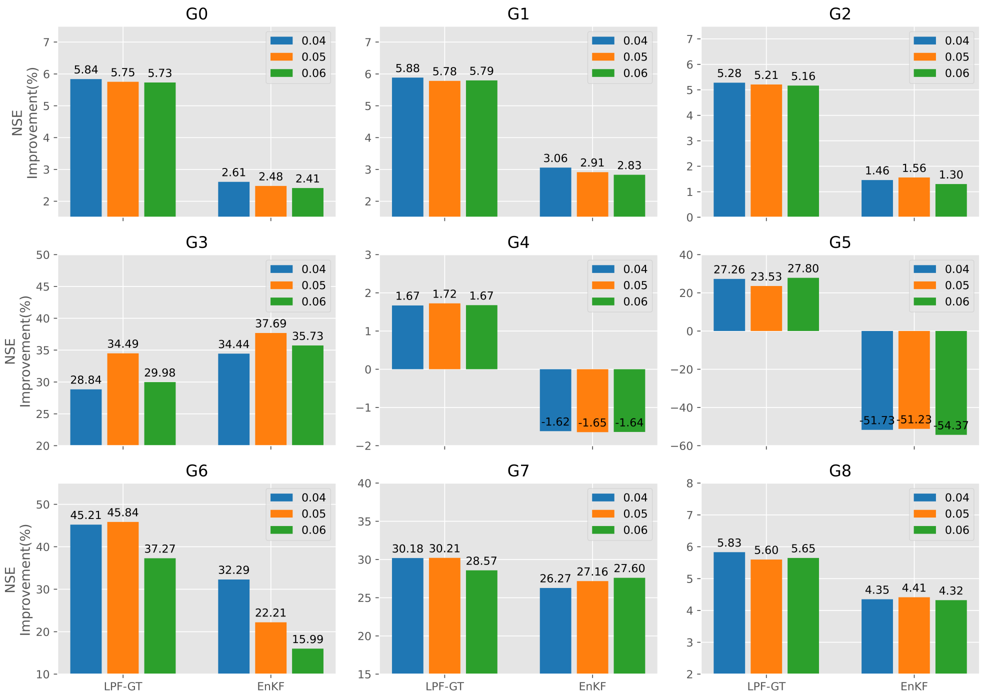

Figure A.9: Improvement percentage in terms of Nash-Sutcliffe index ($NSE$) from DA experiments with different observations' standard deviations using LPF-GT and EnKF with respect to the deterministic run at all validation gauge stations.

# Curriculum Vitæ

## Zhenwu WANG

24-09-1990 Born in Xin Zhou city, Shanxi province, China.

## Education

2009–2013 Undergraduate in Agricultural Water Resources Engineering
Sichuan Agricultural University
*Thesis:* None-cohesive non-uniform sediment scour rate of clear water
*Promotor:* Prof. Lijian Qi

2013–2016 Graduate in Agricultural Water and Soil Engineering
Zhejiang University(ZJU)
*Thesis:* The estimation of reference crop evapotranspiration based on data mining and artificial intelligence
*Promotor:* Prof. Zonglou Guo

2016–2021 PhD. in Data assimilation
Delft University of Technology
*Thesis:* Data assimilation in high dimensional systems using local particle filters
*Promotor:* Prof. dr. ir. Nick van de Giesen
*Copromotor:* dr. ir. Rolf Hut

# List of Publications

## Journal papers

3. **Wang, Z.**, Hut, R., Tangdamrongsub, N., & Van de Giesen, N.(2021) Data assimilation of SMAP soil moisture into the PCR-GLOBWB hydrological model to improve discharge estimates via A Novel Local Particle Filter. *Hydrology and Earth System Sciences* (Submitted).

2. **Wang, Z.**, Hut, R., & Van de Giesen, N.(2021) A Novel Local Particle Filter Based on Gaussian Process Regression for Highly Nonlinear Observation Operator in High-Dimensional Models. *Journal of Advances in Modeling Earth Systems* (In review).

1. **Wang, Z.**, Hut, R., & Van de Giesen, N. (2020). A Local Particle Filter Using Gamma Test Theory for High–Dimensional State Spaces. *Journal of Advances in Modeling Earth Systems*, 12(11), e2020MS002130.

## Conference abstracts

5. **Wang, Z.**, Hut, R., & van de Giesen, N. (2020, May). A Python package for data assimilation in the eWatercycle program-a hydrological framework. In EGU General Assembly Conference Abstracts (p. 11457).

4. **Wang, Z.**, Hut, R., & Van De Giesen, N. (2019, December). A New Localized Particle Filter Based on Gaussian Process Regression for High Non-linear Observation Operators in data assimilation. In AGU Fall Meeting Abstracts (Vol. 2019, pp. NG21B-0911).

3. **Wang, Z.**, Hut, R., & van de Giesen, N. (2019, April). A Localized Particle Filter for High-Dimensional Models. In EGU General Assembly Conference Abstracts (p. 6023).

2. **Wang, Z.**, Hut, R., & van de Giesen, N. (2018, April). A New Local Particle Filter Using Gamma Test Theory for High-Dimensional Models. In EGU General Assembly Conference Abstracts (p. 1619).

1. **Wang, Z.**, Hut, R., & van de Giesen, N. (2017, April). A New Method to Improve Performance of Resampling Process in Particles Filter by Genetic Algorithm and Gamma Test Algorithm. In EGU General Assembly Conference Abstracts (p. 7654).