JAMES | Journal of Advances in Modeling Earth Systems

RESEARCH ARTICLE
10.1029/2023MS003789

Special Section:
Machine learning application to Earth system modeling

A Hybrid Data-Driven and Data Assimilation Method for Spatiotemporal Forecasting: PM2.5 Forecasting in China

Shengjuan Cai[1,2], Fangxin Fang[1,2], Xiao Tang[3], Jiang Zhu[3], and Yanghua Wang[1,2]

[1]Resource Geophysics Academy, Imperial College London, London, UK, [2]Department of Earth Science and Engineering, Imperial College London, London, UK, [3]International Center for Climate and Environment Sciences, Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing, China

Abstract Spatiotemporal forecasting involves generating temporal forecasts for system state variables across spatial regions. Data-driven methods such as Convolutional Long Short-Term Memory (ConvLSTM) are effective in capturing both spatial and temporal correlations, but they suffer from error accumulation and accuracy loss as forecasting time increases due to the nonlinearity and uncertainty in physical processes. To address this issue, we propose to combine data-driven and data assimilation (DA) methods for spatiotemporal forecasting. The accuracy of the data-driven ConvLSTM model can be improved by periodically assimilating real-time observations using the ensemble Kalman filter (EnKF) approach. This proposed hybrid ConvLSTM-EnKF method is demonstrated through PM2.5 forecasting in China, which is a challenging task due to the complexity of topographical and meteorological conditions in the region, the need for high-resolution forecasting over a large study area, and the scarcity of observations. The results show that the ConvLSTM-EnKF method outperforms conventional methods and can provide satisfactory operational PM2.5 forecasts for up to 1 month with spatially averaged RMSE below 20 μg/m^3 and correlation coefficient (R) above 0.8. In addition, the ConvLSTM-EnKF method shows a substantial reduction in CPU time when compared to the commonly used NAQPMS-EnKF method, up to three orders of magnitude. Overall, the use of data-driven models provides efficient forecasts and speeds up DA. This hybrid ConvLSTM-EnKF is a novel operational forecasting technique for spatiotemporal forecasting and is used in real spatiotemporal forecasting for the first time.

Plain Language Summary This study introduces an advanced method (ConvLSTM-EnKF) for PM2.5 forecasting in China, which is a challenging task due to its large area coverage, and complex topographical and meteorological conditions. This innovative approach combines two techniques: one looks at historical data to make forecasts, while the other periodically incorporates new information from observations to improve forecasts over time. This combination significantly improves forecasting accuracy and provides reliable operational PM2.5 forecasts for up to 1 month. Notably, this method is more efficient than traditional approaches. Beyond air pollution, the method holds promise for improving predictions in other areas, including weather, climate, and environmental systems, marking a substantial step forward in our ability to anticipate and understand complex spatiotemporal phenomena.

Key Points:

- A hybrid data-driven (ConvLSTM) and data assimilation (EnKF) method is proposed for accurate and efficient spatiotemporal forecasting. A pre-trained ConvLSTM is used for both the forecasting and assimilation processes, enabling fast online operational forecasting
- The ConvLSTM-EnKF demonstrates high efficiency by reducing CPU time by three orders of magnitude compared to the NAQPMS-EnKF, a widely used air quality forecasting model in China
- The ConvLSTM-EnKF enables online data assimilation (DA) for high-dimensional systems and improves DA accuracy by allowing the use of a large ensemble size

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

F. Fang,
f.fang@imperial.ac.uk

Citation:

Cai, S., Fang, F., Tang, X., Zhu, J., & Wang, Y. (2024). A hybrid data-driven and data assimilation method for spatiotemporal forecasting: PM2.5 forecasting in China. *Journal of Advances in Modeling Earth Systems*, *16*, e2023MS003789. https://doi.org/10.1029/2023MS003789

Received 27 APR 2023
Accepted 14 JAN 2024

Author Contributions:

Conceptualization: Shengjuan Cai, Fangxin Fang

1. Introduction

Spatiotemporal forecasting is crucial for scientific studies and practical applications, such as numerical weather, air quality, and flooding forecasting (Brunner et al., 2020; Cheng et al., 2022; Qi et al., 2019; Ravuri et al., 2021; L. Xu et al., 2021). Accurate and efficient spatiotemporal forecasting can help decision-making and reduce life and economic damage. However, spatiotemporal forecasting is challenging due to several factors, including the inherent complexity of the studied systems, high dimensionality, data sparsity and incompleteness, uncertainties in models and measurements, interactions between variables, and spatial correlations among regions (Chai et al., 2020; Liu et al., 2019).

Machine learning (ML) methods have emerged as a powerful tool for spatiotemporal forecasting, given their ability to handle large data sets efficiently and learn complex input-output relationships from historical data (Kochkov et al., 2021; Liang et al., 2020; Rasp & Thuerey, 2021; Weyn et al., 2019). In particular, recurrent neural networks (RNNs) and RNN-variants, such as Long short-term memory (LSTM), are powerful in exploring temporal dependency in sequential data (Elman, 1990; Hochreiter & Schmidhuber, 1997; Karevan & Suykens, 2020; Li et al., 2017; Yan et al., 2021). By embedding convolutional operators into the LSTM framework, the

Data curation: Xiao Tang, Jiang Zhu
Formal analysis: Shengjuan Cai, Fangxin Fang
Funding acquisition: Yanghua Wang
Methodology: Shengjuan Cai, Fangxin Fang
Project administration: Fangxin Fang, Yanghua Wang
Resources: Fangxin Fang, Jiang Zhu, Yanghua Wang
Software: Xiao Tang, Jiang Zhu
Supervision: Fangxin Fang, Yanghua Wang
Validation: Shengjuan Cai, Xiao Tang
Visualization: Shengjuan Cai
Writing – original draft: Shengjuan Cai
Writing – review & editing: Shengjuan Cai, Fangxin Fang, Yanghua Wang

so-called convolutional LSTM (ConvLSTM) network can incorporate spatial dependence and is widely adopted in spatiotemporal forecasting problems (Alléon et al., 2020; Kim et al., 2017; Shi et al., 2015). However, most studies are limited in short-term forecasting due to the increased uncertainty associated with longer time horizons and the potential impacts of unforeseen events. To tackle this challenge, we propose to incorporate real-time observations through DA to periodically refine forecasting trajectory and minimize forecasting uncertainty.

DA can improve forecasting accuracy by reducing uncertainty through the fusion of observations and model simulations (Park & Xu, 2022). DA is often used in forecasting problems of chaotic systems, like weather and air quality forecasting, where even a small disturbance of the initial state can lead to huge differences in long-term forecasting results (Lorenz, 1963; Schemm et al., 2023). By periodically incorporating information from newly received observations, we can have an optimal initial state enabling more accurate and reliable forecasting results (Ghil & Malanotte-Rizzoli, 1991). DA has been a key component in atmospheric chemistry models for several decades (Bocquet et al., 2015; Elbern & Schmidt, 2001; Sandu & Chai, 2011). The effective utilization of DA techniques has notably improved the accuracy of air pollutant forecasts, particularly for PM2.5 (Particle Matter with a diameter equal to or less than 2.5 microns) (Dash et al., 2023; Peng et al., 2017). DA incorporates real-time observations, such as measurements of pollutant concentrations from ground-based sensors, satellite data, and other sources, into the model. By assimilating this observational data, DA helps correct any discrepancies between model forecasts and actual air quality observations.

One popular DA method, the ensemble Kalman filter (EnKF), operates by generating an ensemble of model state samples, representing the uncertainty in the initial condition (Hakim et al., 2016; Zhu et al., 2019). These samples are then projected forward to generate a set of forecasts that encompass the possible future states. EnKF assimilates observations, when available, to update the ensemble and produce a more accurate estimate of the current system state. EnKF has the advantage of being computationally efficient, easy to implement and adaptable to different models and systems compared to other DA methods such as four-dimensional variational DA (4D-Var). However, EnKF is sensitive to the used ensemble size (Lorenc, 2003). A small ensemble size can result in inaccurate assimilation results, while a large ensemble size can be computationally expensive (Evensen, 1994; Evensen et al., 2022). Typically, $O(10^2)$ ensemble members are used in conventional EnKF procedures with physics-based forecasting models due to limited computational resources (Houtekamer et al., 2014; Houtekamer & Zhang, 2016; Leutbecher, 2019).

Recent studies are exploring the combination of ML and DA methods to reduce the accumulation of forecast uncertainty and address the challenges posed by intensive computation in conventional physics-based models. Penny et al. (2022) implemented a pre-trained RNN as a surrogate model for prediction in a sequential DA cycle and tested the integrated RNN-DA approach in the Lorenz 96 system. Chattopadhyay et al. (2022) combined a data-driven weather prediction (DDWP) model with DA and demonstrated the proposed DDWP + DA approach on Z500 (geopotential height at 500 hPa) forecasting. These studies demonstrated that the integrated approaches can correct forecast trajectories and represented an important step toward combining a purely data-driven model with DA for operational spatiotemporal forecasting. However, these studies are only demonstrated in toy and simple systems with low dimensionality and synthetic observations. Pawar and San (2022) adopted a model order reduction algorithm to reduce the dimensionality of the studied system. They proposed to train an LSTM network as a surrogate model and perform DA in the reduced-order space. However, the reduced model sacrifices accuracy for efficiency and may not be reliable for highly nonlinear systems. Meanwhile, the spatial correlations of adjacent regions and complex intercorrelations between different state variables in the studies system are not exploited. Additionally, the scarcity and sparseness of observations in realistic scenarios can also pose significant challenges in DA.

In this work, we propose a ConvLSTM-EnKF method for operational spatiotemporal forecasting, in which a ConvLSTM model is trained and used for iterative multi-step forecasting and EnKF is used to correct forecasts by assimilating real-time observations into the ConvLSTM forecasts. The ConvLSTM model is trained on a reanalysis data set produced by assimilating observations into physical simulation results. This is a hybrid data-driven and DA approach that combines the strengths of ML algorithms, real-time observations, and physics-based models. Our contributions focus on (a) ML-based operational spatiotemporal forecasting with high efficiency and accuracy; (b) online DA with high spatial resolution; (c) the use of sparse observations in DA; (d) the incorporation of complex topographical and meteorological features on ML-based forecasting; (e) the application of the proposed ConvLSTM-EnKF method in PM2.5 forecasting over China.

This paper is structured as follows: In Section 2, we propose an operational spatiotemporal forecasting method by combining ConvLSTM and EnKF. This section provides the governing equations for spatiotemporal forecasting, explanations of ConvLSTM and EnKF, and the structure of the proposed ConvLSTM-EnKF method. Section 3 describes the application of the proposed ConvLSTM-EnKF method in operational PM2.5 forecasting, including the study area, the data sets used, and the details of model construction, training, and validation. In Section 4, we conduct a comprehensive evaluation and analysis of both the ConvLSTM and ConvLSTM-EnKF performances in PM2.5 forecasting. Finally, Section 5 presents conclusions drawn from the preceding results.

2. An Efficient ConvLSTM and EnKF DA Framework

2.1. ML-Based Modeling for Nonlinear Forecasting Problems

Complex physical processes can be governed by partial differential equations (PDEs), such as the continuity equation, Navier-Stokes (NS) equation, thermal diffusion equation, and wave equation. The governing equations for these problems can be generally expressed as:

$$\frac{d\boldsymbol{\varphi}}{dt} = g(\boldsymbol{\varphi}, \mathbf{u}, t), \tag{1}$$

where $\boldsymbol{\varphi}$ denotes the state variable to be forecasted in a spatial region, such as velocity, pressure, temperature, and concentrations. $\mathbf{u}$ represents the model inputs, including physical parameters, and initial and boundary conditions. t is time and g denotes a nonlinear function.

In general, forecasting problems involve a nonlinear relationship between the state variable at the current time level t and the next time level $t+1$, given the model input $\mathbf{u}_{t+1}$ at time $t+1$. This relationship can be expressed as:

$$\boldsymbol{\varphi}_{t+1} = M\left(\boldsymbol{\varphi}_t, \mathbf{u}_{t+1}\right), \tag{2}$$

where M denotes a nonlinear forecast operator.

It is often not possible to find analytical solutions for the governing equations in a complex system. Therefore, numerical methods are commonly utilized to solve the associated PDEs. However, widely used numerical methods, like the finite difference method (FDM), finite element method (FEM), and finite volume method (FVM), can introduce numerical errors and can be computationally expensive, particularly for large-scale problems. Recent advancements have demonstrated that ML methods can offer a feasible and efficient way to perform spatiotemporal forecasting by leveraging the nonlinear relationships learned from data. We compared data-driven and physics-based models for spatiotemporal forecasting from the perspective of accuracy and efficiency. Detailed comparative analysis between data-driven and physics-based methods can be found in Appendix in Supporting Information S1, Text S1 in Supporting Information S1.

The ConvLSTM model adopted in this study is an effective approach for spatiotemporal forecasting, which captures both spatial and temporal dependencies in the data by utilizing convolution operators in both the input and hidden states of the LSTM cells. For more details about the ConvLSTM model, please refer to the Appendix in Supporting Information S1. We train a ConvLSTM model for hourly forecasting and achieve multi-hour-ahead forecasting by iteratively using the trained hourly forecasting model. In this iterative forecasting procedure, each prior forecast is employed as input for sequential forecasting. We chose the iterative strategy because the required hourly forecasting model is easy to train and it can generate forecasts of arbitrary lengths (Shi & Yeung, 2018). More information on iterative multi-step-ahead (IMS) forecasting strategy can be found in the Appendix in Supporting Information S1, Text S2.2 in Supporting Information S1.

2.2. Coupling of EnKF With ConvLSTM

Real-time DA is crucial for improving operational forecasting accuracy. A widely adopted DA method, known as EnKF, leverages an ensemble of model simulations and incorporates real-time observations to optimally estimate the state of a system. EnKF is relatively easy to implement and model-independent, making it a preferred option over other DA methods like 4D-Var. However, EnKF requires a large ensemble size to accurately represent system uncertainty, which can be computationally demanding, particularly for large-scale systems, and often

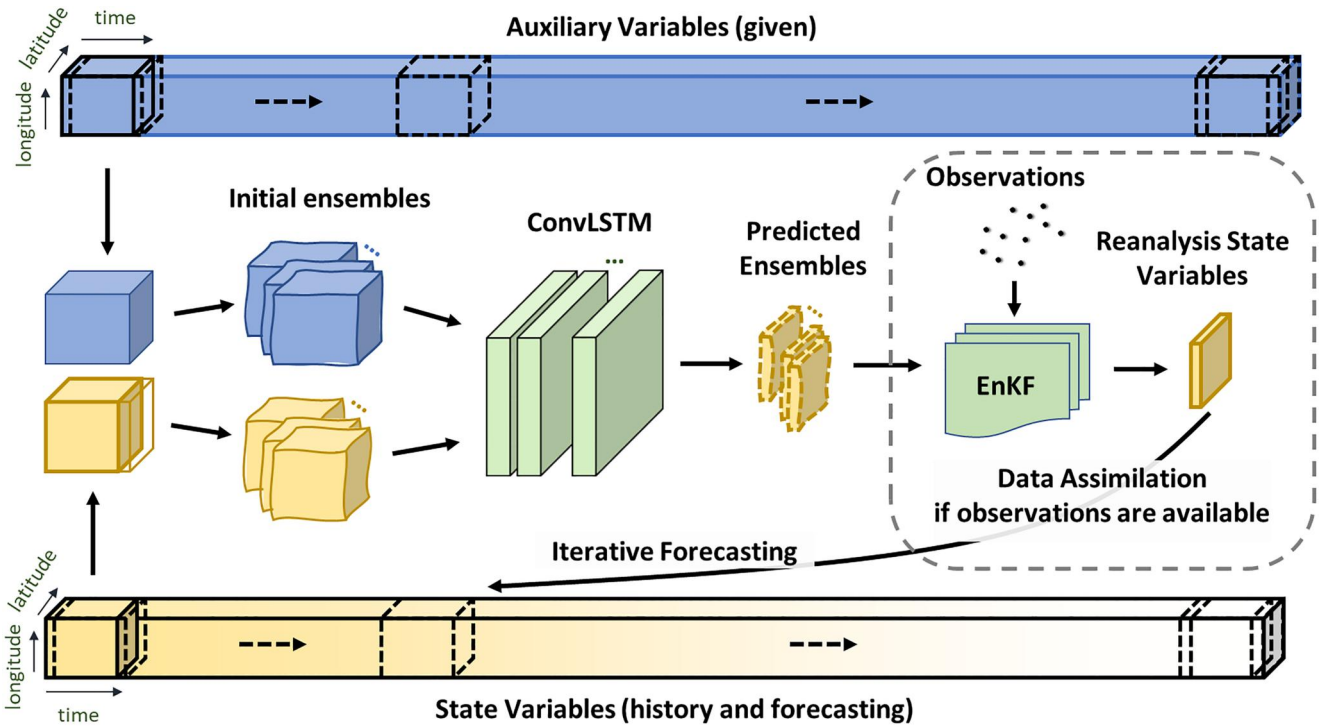

Figure 1. ConvLSTM-EnKF method for operational spatiotemporal forecasting. Auxiliary variables are available or can be obtained from other systems. State variables are what we are going to forecast. We first train a ConvLSTM model for hourly forecasting and further perform multi-hour-ahead forecasting by adopting the IMS strategy. Ensemble Kalman filter (EnKF) is used to assibilate newly received observations to reduce forecast uncertainty and generate optimal reanalysis state variables.

requires parallel computing resources for practical runtime performance. To tackle this limitation, we propose a hybrid method that employs a data-driven ConvLSTM model as the forecasting model in EnKF. This ConLSTM-EnKF method enables the use of a large ensemble size, improving both the efficiency and accuracy of DA. The combination of data-driven and DA methods can further benefit online DA and operational spatiotemporal forecasting. More information on EnKF can be found inAppendix in Supporting Information S1, Text S3 in Supporting Information S1.

The nonlinear forecast operator M in Equation 2 can be either physics-based or data-driven models. Here, a trained ConvLSTM model M_{ConvLSTM} is used for forecasting in the proposed ConvLSTM-EnKF method. The forecast process is expressed as:

$$\boldsymbol{\varphi}_{t+1} = M_{\text{ConvLSTM}}\left(\boldsymbol{\varphi}_t, \mathbf{u}_{t+1}\right) + \boldsymbol{\varepsilon}_{t+1}, \tag{3}$$

where $\boldsymbol{\varepsilon}_{t+1}$ is the model error. EnKF is used to reduce $\boldsymbol{\varepsilon}_{t+1}$ by incorporating observations into the ML-based model M_{ConvLSTM} simulations. There are three steps in the proposed hybrid ConvLSTM-EnKF method (Figure 1):

1. Train a ConvLSTM model for spatiotemporal forecasting.
2. Estimate the system state by conducting ensemble forecasting using the trained model.
3. Update the estimated state by assimilating real-time observations into the forecasted ensemble members using EnKF when observations are available.

3. A Case Study of PM2.5 Forecasting in China

PM2.5 forecasting is crucial for anticipating and preparing for potential health and environmental impacts. The complex behavior of air pollutants, such as PM2.5, can be described through the interaction of various processes, including advection, diffusion, chemical reactions, emissions, and deposition. These complex processes are often expressed mathematically through a set of coupled PDEs. Here, we employed an advection-diffusion equation with a source term to broadly encapsulate the complex atmospheric processes relevant to PM2.5 forecasting:

$$\frac{\partial \varphi}{\partial t} + \nabla \cdot (\mathbf{v}\varphi) = \nabla \cdot (D\nabla\varphi) + S_{\varphi}, \tag{4}$$

where φ represents the PM2.5 concentration; $\mathbf{v}$ is the wind velocity vector; D is the diffusion coefficient; S_{φ} is the source term that encompasses PM2.5 emission, deposition, and related chemical reactions, each with its specific mathematical formulations designed to the underlying processes.

We evaluate the capability of the proposed ConvLSTM-EnKF method through a case study on operational spatiotemporal PM2.5 forecasting in China. This task is challenging since (a) it covers the vast territory of China with a variety of terrain, economic activities and meteorological conditions that exert significant impacts on PM2.5 concentrations (An et al., 2019; Hong et al., 2019; Zhang et al., 2019); (b) it involves a high-resolution (15 × 15 km in spatial and 1 hr in temporal) data set which is necessary for capturing complex atmospheric processes at different scales and accurately predicting air pollution patterns, but can lead to a dramatic increase in computational demands and make it difficult to perform online DA (Kong et al., 2021); (c) it involves the assimilation of sparsely located observations, which poses difficulties in generating reanalysis results with sufficient accuracy. In this study, we tackle the first challenge mentioned by incorporating topographical and meteorological features into the data-driven forecasting model, aligning with the approach of physics-based models. To alleviate the computational cost-related difficulties, we adopt a ConvLSTM model as a surrogate for physics-based models. This enhanced forecasting efficiency achieved by the ConvLSTM model allows the utilization of a large ensemble size in EnKF for DA, ultimately leading to improved analysis accuracy.

3.1. High-Resolution Reanalysis Data and Sparse Observations

Reanalysis PM2.5 Concentrations are from the Chines air quality reanalysis (CAQRA) data set. This data set is obtained by assimilating the surface observations from the sparse monitoring stations into the physical simulation results obtained by the Nested Air Quality Prediction Modeling Systems (NAQPMS) using EnKF (Kong et al., 2021). As a result, the CAQRA data set inherently encompasses atmospheric processes such as advection, diffusion, chemical reactions, emissions, and deposition from the NAQPMS simulations. This data set contains six conventional air pollutants (PM2.5, PM10, SO2, NO2, CO, and O3) in China for the period 2013–2019 at high spatial (15 km × 15 km) and temporal (1 hr) resolutions. This reanalysis data set is used as the ground truth for training and validating the ConvLSTM model as it optimally combines the surface observations and physical modeling results through advanced assimilation techniques.

Meteorology and topography data are integrated as inputs during both the training and forecasting stages of the ConvLSTM model. Meteorological variables, including the U- and V-components of wind, temperature, relative humidity, and surface pressure are obtained through meteorology simulations using the Weather Research and Forecasting (WRF) model (Skamarock et al., 2008). Each simulation involves a continuous 36-hr run of the WRF model, with the first 12 hr serving as a spin-up phase and the subsequent 24 hr providing the meteorological inputs for the ConvLSTM model. For each simulation, the initial and boundary conditions are derived from the Climate Forecast System Reanalysis (CFSR) data set (Saha et al., 2010). Altitude is obtained from a Digital Elevation Model (DEM).

Sparse surface PM2.5 observations are collected from 1,436 sparsely located air quality monitoring stations across China. These stations are unevenly distributed, with a majority located in eastern China and a few stations in the western regions. The uneven distribution of observations poses challenges for DA in PM2.5 forecasting. The specific locations of all monitoring stations can be found in Appendix in Supporting Information S1, Figure S1 in Supporting Information S1.

3.2. The ConvLSTM Model for Hourly PM2.5 Forecasting

In this study, a 3-layer ConvLSTM model is trained for hourly PM2.5 forecasting. Details of model construction, training, and validation can be found in Appendix in Supporting Information S1, Text S2.3 in Supporting Information S1, This ConvLSTM model is initialized daily from reanalysis data generated by EnKF at 00:00 UTC+08:00 in March, June, September, and December 2018. The IMS strategy is used to forecast PM2.5 for the following 24 hr.

The hourly forecasting model can be represented as:

$$\mathbf{O}_t = M_{\text{ConvLSTM}}(\mathbf{I}_t), \tag{5}$$

where $\mathbf{I}_t$ and $\mathbf{O}_t$ represent the input and output of the ConvLSTM model M_{ConvLSTM} at time t, respectively.

$$\mathbf{I}_t : \begin{cases} \left\{\mathbf{u}^k_{t-m+2:t+1}\right\} \in \mathbb{R}^{m\times N}, k = 1,2,\cdots K \\ \left\{\boldsymbol{\varphi}_{t-m+1:t}\right\} \in \mathbb{R}^{m\times N}, \end{cases} \tag{6}$$

where input $\mathbf{I}_t$ includes the auxiliary variables $\mathbf{u}$ (here topographical and meteorological variables) and historical PM2.5 concentrations $\boldsymbol{\varphi}$; m is the time-lag length ($m = 1$ in this study); K is the number of auxiliary variables, $\mathbf{u}^k \in \mathbb{R}^N$ represents the kth auxiliary variable with N nodes in the study domain. $\mathbf{u}_{t+1}$ at time $t + 1$ is either a time-invariant variable (i.e., altitude) or a time-variant variable forecasted from a related model (i.e., meteorological variables forecasted from the WRF model). The output $\mathbf{O}_t$ is the forecasted PM2.5 concentrations $\widetilde{\boldsymbol{\varphi}}_{t+1}$ with lead time of 1 hr:

$$\mathbf{O}_t : \widetilde{\boldsymbol{\varphi}}_{t+1} \in \mathbb{R}^N. \tag{7}$$

In the training and validation processes, reanalysis data from 2013 to 2017 is utilized. 90% of this data is randomly allocated for training, while the remaining 10% is used for validation. The training and validation loss can be found in Appendix in Supporting Information S1, Figure S4 in Supporting Information S1. To ensure independence, all samples for training and validation are shaffled before being fed into the model. Input variables of the ConvLSTM model include historical PM2.5 concentration, U- and V-components of wind fields, temperature, relative humidity, and altitude (Figure 2). These input variables are optimized based on correlation analysis among all available variables. More details on the correlation analysis and related discussions can be found in Appendix in Supporting Information S1 Text S4 in Supporting Information S1. The ConvLSTM model, being data-driven, derives valuable insights from its training data, specifically the CAQRA data set. This data set is created by skillfully integrating NAQPMS simulations and surface observations through DA. This integration provides the model with a comprehensive understanding of complex atmospheric processes, including advection, diffusion, chemical reactions, and emissions.

3.3. The ConvLSTM-EnKF Model for Operational PM2.5 Forecasting

Observing System Simulation Experiment (OSSE): We conduct an OSSE to assess the performance of the proposed ConvLSTM-EnKF method in hourly operational PM2.5 forecasting and analyze the impact of different DA frequencies on the results. A set of simulated observations are assimilated into the ConvLSTM forecasts every 6, 12, and 24 hr. The synthetic observations are created by down-sampling reanalysis PM2.5 concentrations with a spatial resolution of 150 × 150 km, where random noise, equivalent to 10% of the original value, is added to simulate measurement errors. These synthetic observations comprise only 1% of the reanalysis data set, representing the scarcity of real-world observations.

Real-world Scenario with Sparse Observations: We further apply the ConvLSTM-EnKF method in a real-world scenario, where surface PM2.5 observations are collected from 1,436 sparsely located air quality monitoring stations in China. DA is performed every 6 hr using EnKF with a configuration of 100 ensemble members and a localization radius of 1,000 km. The proposed ConvLSTM-EnKF method for operational PM2.5 forecasting is illustrated in Figure 2.

4. Results and Discussion

4.1. ConvLSTM Performance in PM2.5 Forecasting

The accuracy of PM2.5 forecasting depends on several factors, including data availability and quality, model effectiveness, and the accuracy of meteorological forecasts. Recently developed data-driven models can achieve reliable PM2.5 forecasting. For example, Niu et al. (2023) proposed an Informer-based spatiotemporal predictor, demonstrating good performance in hourly PM2.5 forecasting at 35 monitoring stations in Beijing. Teng et al. (2023) employed a graph deep neural network to capture the spatiotemporal correlations among

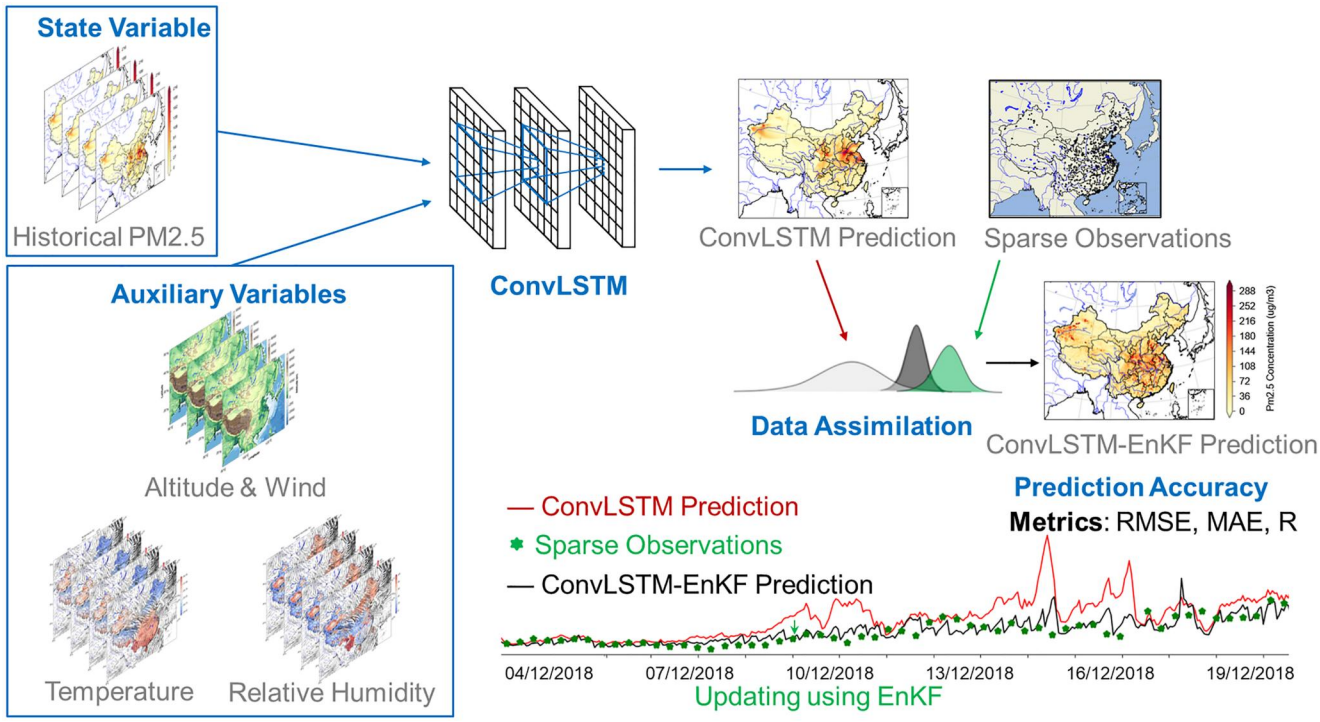

Figure 2. Illustration of the proposed hybrid ConvLSTM-EnKF method for hourly operational PM2.5 forecasting. In this case study, PM2.5 concentrations serve as the state variable, and topographical and meteorological fields, including U- and V-components of the wind field, temperature, relative humidity, and altitude, are used as auxiliary variables. Observations are collected from 1,436 sparsely located air quality monitoring stations in China. Forecasting is conducted using a trained three-layer ConvLSTM model while EnKF is used to improve the forecast accuracy by assimilating the observations into the ConvLSTM forecasts.

neighborhood monitoring stations for PM2.5 forecasting in the Beijing-Tianjin-Hebei region. However, most studies only demonstrated at specific stations, cities, or regions, PM2.5 forecasting over a large area, like the entire China, remains a challenging task.

In this section, we demonstrate the ability of the trained ConvLSTM model in hourly PM2.5 forecasting through a comparison with surface monitoring data and reanalysis data (served as reference data) in a high spatial resolution of 15 km × 15 km. We evaluate the performance of the trained ConvLSTM model in hourly PM2.5 forecasting with a lead time of 24 hr during December 2018, a winter season characterized by severe PM2.5 pollution. The model is initialized daily from reanalysis data at 00:00 UTC+08:00. Subsequently, PM2.5 concentrations for the following 24 hr were forecasted using the IMS strategy. The inputs for the ConvLSTM are historical PM2.5 concentrations, key meteorological (wind speed, temperature, relative humidity) and topographical (altitude) factors, which largely impact PM2.5 distribution and concentration in the atmosphere (Song & Shao, 2023; Xu et al., 2018). Detailed correlation analysis among all available factors can be found in Appendix in Supporting Information S1, Text S4 in Supporting Information S1.

Our results show that the trained ConvLSTM model can provide efficient and accurate hourly PM2.5 forecasting with a lead time of 24 hr in December 2018. The forecasted results are compared with reanalysis PM2.5 concentrations and actual observations in four cities in China: Beijing, Shanghai, Guangzhou, and Chengdu (Figure 3). Our analysis reveals that the ConvLSTM forecasts exhibit a high degree of alignment with the reanalysis data and actual observations. The temporally averaged RMSE values between forecasts and observations in four cities are 24.82, 28.81, 13.51, and 31.84 $\mu g/m^3$, respectively. The trained ConvLSTM effectively captures the temporal variation in PM2.5 concentrations throughout the entire month, emphasizing its accuracy in forecasting pollution levels under severe winter conditions. Satisfactory 24-hr-ahead forecasts for the other 3 months (March, June, and September) in 2018 are presented in Appendix in Supporting Information S1, Figure S5–S7 in Supporting Information S1.

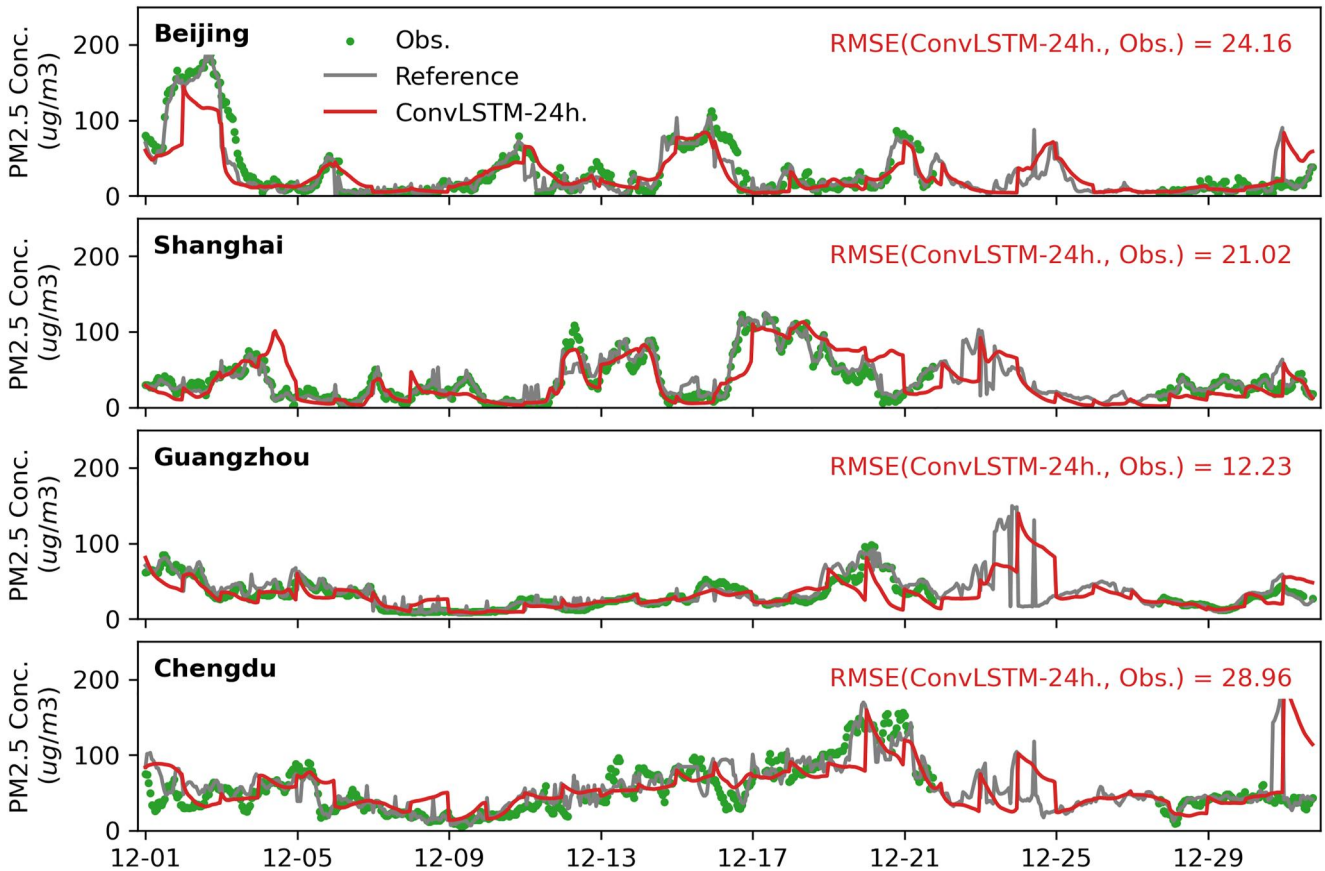

Figure 3. 24-hour-ahead PM2.5 forecasts in 4 cities (Shanghai, Beijing, Chengdu, and Guangzhou) in China during December 2018. Forecasts are initiated daily at 0:00 UTC+08:00, employing reanalysis PM2.5 concentration as the initial condition. RMSE(ConvLSTM-24hr, Obs.) denotes the temporally averaged RMSE between the 24-hr-ahead ConvLSTM forecasts and observations for each city.

We also assessed how well the ConvLSTM model performs by comparing it to the NAQPMS, a widely used physics-based model for air quality forecasting in China, specifically for hourly PM2.5 forecasting with a lead time of 1 month. Our findings reveal that the ConvLSTM model outperforms the NAQPMS (Appendix in Supporting Information S1, Figure S8 in Supporting Information S1). For an extensive examination and comparison between data-driven and physics-based models, please refer to the detailed analysis provided in the Appendix in Supporting Information S1, Text S1 in Supporting Information S1.

Furthermore, our results indicate that the trained ConvLSTM model can capture the spatial distribution of the PM2.5 concentration (Figure 4). The absolute forecast error at 06:00 UTC+08:00, 19 December 2018, is smaller than 10 $\mu g/m^3$ in most areas. The largest forecast error with an absolute value of 40 $\mu g/m^3$ was observed in the Jing-Jin-Ji region characterized by a high population density and developed industries. The spatially averaged RMSE and R between the forecasts and reference are 8.88 $\mu g/m^3$ and 0.95, respectively. Additionally, local correlations between predictions and references are above 0.80 in all areas (Appendix in Supporting Information S1, Figure S10 in Supporting Information S1). The mathematical definition of local correlation can be found in Appendix in Supporting Information S1, Text S5 in Supporting Information S1.

However, we recognize that the ConvLSTM model also suffers from accumulation and accuracy loss over forecasting time, just like other physics-based and data-driven models for PM2.5 forecasting, which is inherent in chaotic systems (Lorenz, 1972). The spatially averaged RMSE increases to 14.26 $\mu g/m^3$, the spatially averaged R decreases to 0.90 at 12:00 UTC+08:00; the spatially averaged RMSE further increases to 14.79 $\mu g/m^3$, the spatially averaged R decreases to 0.84 at 18:00 UTC+08:00. To mitigate the accuracy loss stemming from errors in both the initial condition and forecasting model, we propose the ConvLSTM-EnKF method in Section 3.3 for

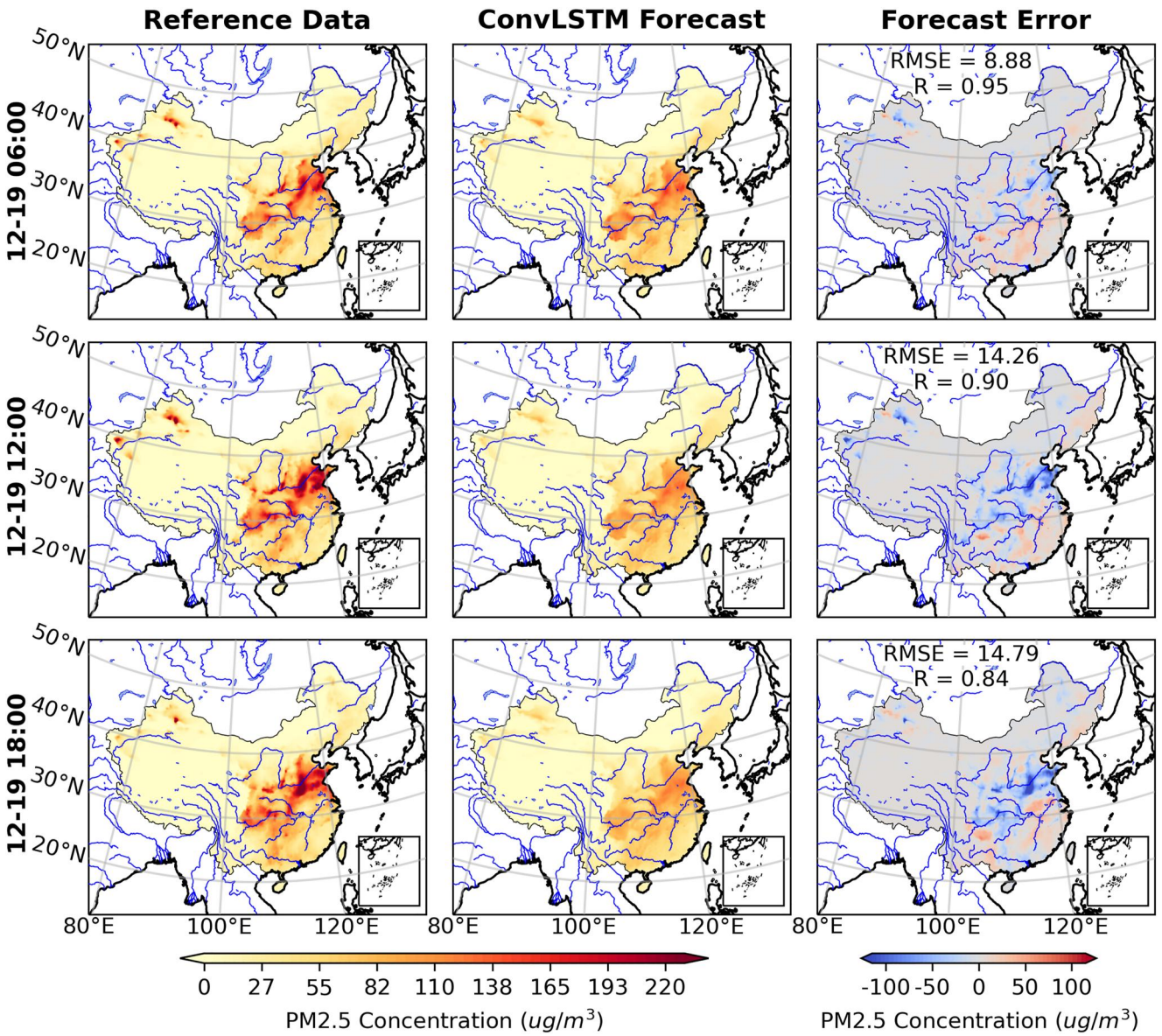

Figure 4. Comparison of the PM2.5 concentrations obtained from the reanalysis data set (left), hourly ConvLSTM forecasts with a lead time of 24 hr (middle), and differences between them (right) at 06:00, 12:00, and 18:00 UTC+08:00 on 19 December 2018.

operational PM2.5 forecasting. This method allows for periodic refinement of the initial condition by assimilating observations into the ConvLSTM forecasts, ensuring reliable and consistent forecasts.

4.2. ConvLSTM-EnKF Performance in Operational PM2.5 Forecasting

To improve the accuracy of operational spatiotemporal forecasting, we integrate the EnKF into the ConvLSTM model, resulting in the ConvLSTM-ENKF method. This approach allows for online DA at a high spatial resolution, using either uniformly located synthetic observations or sparsely located monitoring observations. The forecasting trajectory is periodically corrected with a selected DA frequency. We demonstrate the effectiveness of this method in improving forecasting accuracy through the following two case studies.

4.2.1. DA With Synthetic Observations

To evaluate the effectiveness of the proposed ConvLSTM-EnKF method, we conduct operational PM2.5 forecasting with a set of synthetic observations. These synthetic observations are created by down-sampling the

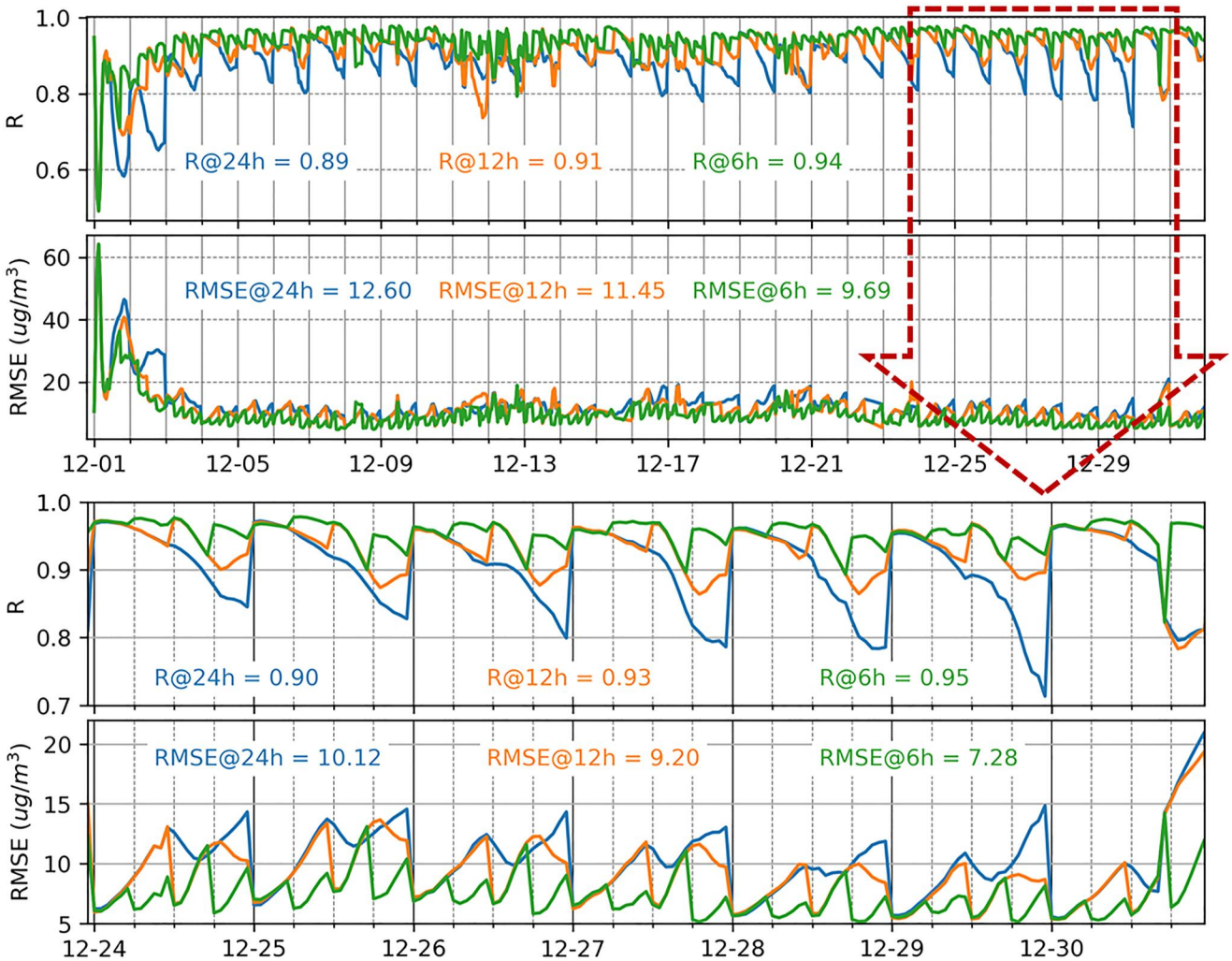

Figure 5. Evaluating ConvLSTM-EnKF operational forecasting accuracy at different DA frequencies (6, 12, and 24 hr) in December 2018. DA is conducted using EnKF with synthetic observations and 100 ensemble members. The blue, yellow, and green curves represent the spatially averaged R and RMSE of PM2.5 concentration between reference data and ConvLSTM-EnKF forecasts at each hour. The labels 'R@Xh' and 'RMSE@Xh' denote the spatiotemporally averaged R and RMSE with a DA frequency of X hours over China throughout the entire month.

reference data with a spatial resolution of 150 × 150 km and introducing random noise at a level corresponding to 10% of the original values. Using the trained ConvLSTM model, we conduct IMS forecasting and incorporate the observations using EnKF at the end of each DA cycle. By employing synthetic observations, we can achieve the highest attainable accuracy in operational PM2.5 forecasting. We systematically analyze the effects of different DA frequencies on the operational forecasting performance. This analysis can help in identifying the optimal DA frequency for operational forecasting with actual observations. In this study, we adopt three DA frequencies (6, 12, and 24 hr) to facilitate the periodic incorporation of observations.

We evaluated the spatiotemporal PM2.5 forecasting results with different DA frequencies by calculating the RMSE and R between the forecasts and reference data (Figure 5). We observe an increase in RMSE and a decrease in R within each DA cycle. At the end of each DA cycle where DA is conducted, the forecasting accuracy is notably improved by incorporating information from observations. The specific values for RMSE reduction and R increase at the end of each DA circle can be slightly different due to fluctuations in weather conditions. The updated result is used as the initial condition for the next DA cycle. In the studied month, the averaged RMSE values between reanalysis data and operational PM2.5 forecasts with DA frequencies of 6, 12, and 24 hr are 12.60, 11.45, and 9.69 μg/m^3, respectively. The corresponding R values are 0.89, 0.91, and 0.94, respectively. The

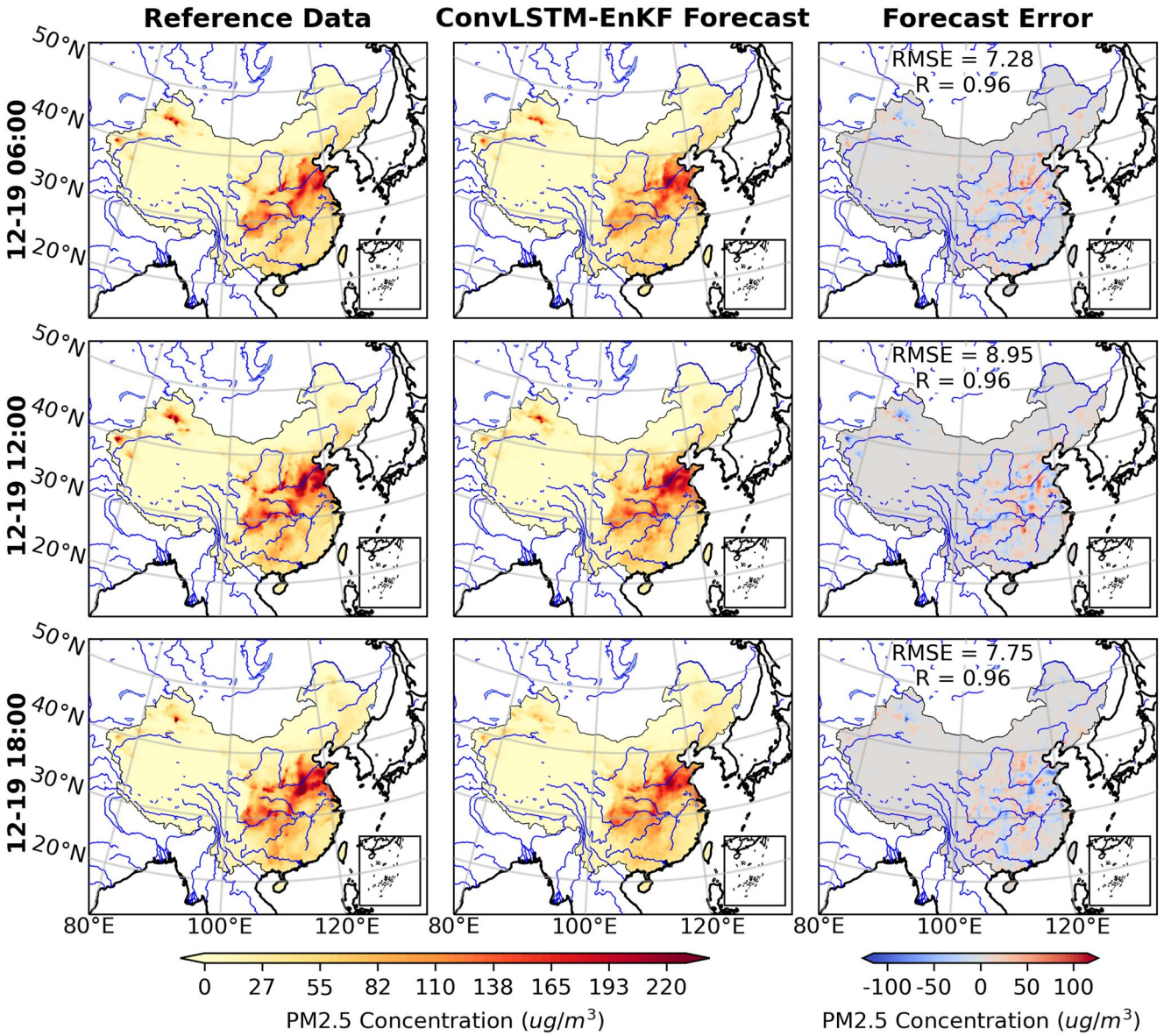

Figure 6. Comparison of the spatial distribution of PM2.5 concentration obtained from the reanalysis data set (left), ConvLSTM-EnKF forecasts (middle), and differences between them (right) at 06:00, 12:00, and 18:00 UTC+08:00 on 19 December 2018. DA frequency is opted as 6 hr for this experiment.

comparison of DA results at different frequencies indicates that shorter DA cycles can enhance operational forecasting accuracy.

The use of the proposed ConvLSTM-EnKF method in operational PM2.5 forecasting can maintain accurate forecasting results over a long period. When the DA frequency is set to 6 and 12 hr, the RMSE maintains below 20.0 μg/m^3 and R above 0.80 throughout the studied month, except for December 1–2, 2018. The abnormally large errors observed during this period are due to unforeseen high PM2.5 concentrations in northwest China presented in the reanalysis data set. Compared with the reference data, we see that the ConvLSTM-EnKF method can capture spatial distributions of PM2.5 concentrations with a small predictive error of 10 μg/m^3 in most areas at 06:00, 12:00, and 18:00 UTC+08:00 on 19 December 2018 (Figure 6). The spatial averaged RMSE values at these times are 7.28, 8.95, and 7.75 μg/m^3, respectively. The corresponding R values are 0.96, 0.96, and 0.96, respectively. The synergy of ConvLSTM and EnKF enables operational spatiotemporal PM2.5 forecasting with high accuracy and stability.

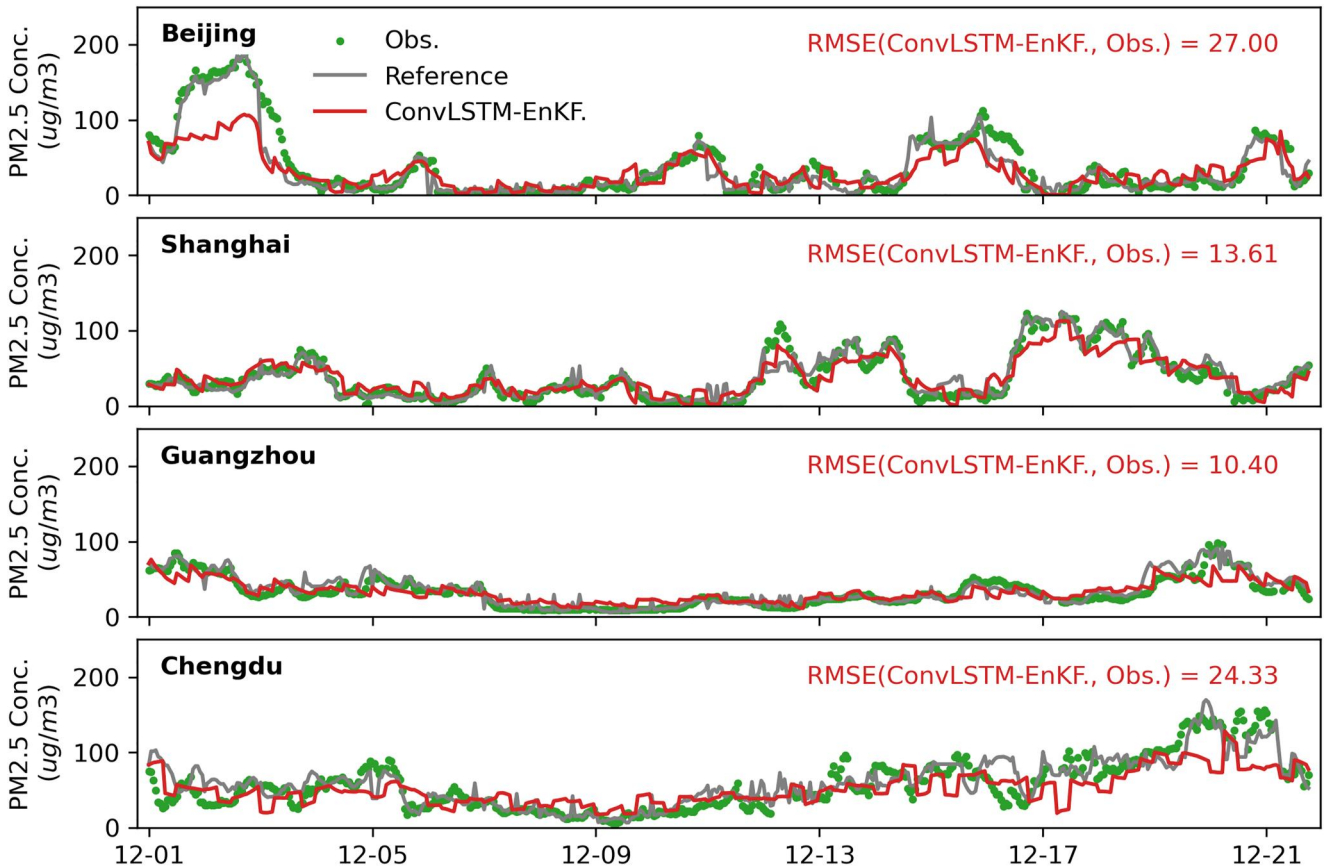

Figure 7. Operational PM2.5 forecasts using the proposed ConvLSTM-EnKF method in cities (Beijing, Shanghai, Chengdu, and Guangzhou) in China during 1–21 December 2018. Data assimilation (DA) is conducted with 100 ensemble members every 6 hr. RMSE(ConvLSTM-EnKF., Obs.) denotes the temporally averaged RMSE between the ConvLSTM-EnKF forecasts and observations for each city.

Hence, in practical scenarios with limited and sparse observations, as discussed in Section 4.2.2, we choose a DA frequency of 6 hr. This decision aims to strike a balance between accuracy and computational efficiency in our forecasting approach.

4.2.2. DA With Sparse Observations

In real-world scenarios, the availability of observations is restricted to a small number of sparsely distributed monitoring stations, which has posed a significant obstacle to achieving reliable DA. To demonstrate the effectiveness of the proposed ConvLSTM-EnKF method in realistic operational PM2.5 forecasting, we use observations from 1,436 sparsely distributed monitoring stations across China during the DA process. DA is conducted every 6 hr using 100 ensemble members. To avoid impractical long-distance correlations, we apply localization with a radius of 500 km at the DA step.

We compare ConvLSTM-EnKF forecasts with actual observations and reanalysis PM2.5 concentrations (reference) from the CAQRA data set in four cities (Beijing, Shanghai, Chengdu, and Guangzhou) in China (Figure 7). Despite differences in economic, topographical, and meteorological conditions contributing to different scales and temporal fluctuations in PM2.5 concentrations, the ConvLSTM-EnKF forecasts align well with the reference and observations for most of the forecasting period (1–21 December 2018). Discrepancies are noted only for high PM2.5 concentrations in Beijing between December 1–3, and in Chengdu between December 19–21. Operational PM2.5 forecasting is halted on December 22 due to the unavailability of observations between December 22–27 (Figure 3). Furthermore, to provide a comprehensive view of the operational PM2.5 forecasting results, satisfactory outcomes for the other 3 months (March, June, and September 2018) are presented in Appendix in Supporting Information S1, Figure S11–S13 in Supporting Information S1.

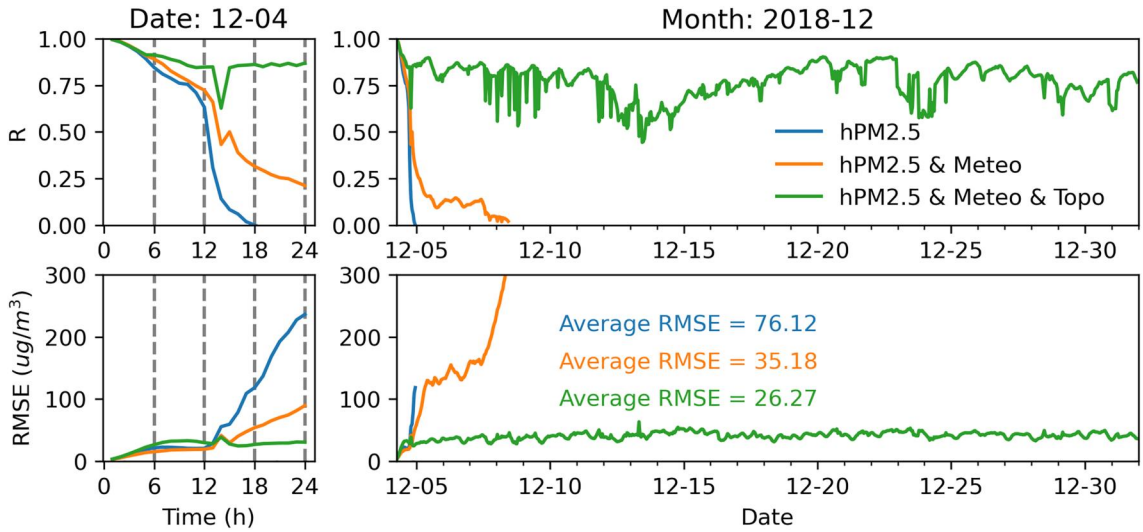

Figure 8. Error analysis and impact of meteorological and topographical features on data-driven PM2.5 forecasting. The R and RMSE values between the ConvLSTM forecasts and reference data on 4 December 2018 (left) and from December 4 to 29, 2018 (right).

4.3. Dependency of PM2.5 Concentration on Meteo- and Topo-Features

PM2.5 forecasting relies on several factors, such as topographical and meteorological conditions. To evaluate the impact of these factors on PM2.5 forecasting, three case studies have been undertaken with different input feature combinations. The key meteorological features considered in the inputs are the U- and V-components of the wind field, temperature, and relative humidity. We employ the spatially averaged RMSE and *R* metrics to quantify the performance of the trained ConvLSTM in these three cases, respectively (Figure 8).

The forecasting method that combines both historical PM2.5 concentrations and meteorological features has demonstrated better performance compared to the method that relies solely on historical PM2.5 concentrations. The latter can only provide accurate forecasts for a single day before the *R* between the forecast and reference drops to zero, while the former can provide accurate forecasts for over 4 days. Moreover, incorporating topographical features into the forecasting model further improves its performance, with a consistently high *R* (above 0.5) and low RMSE (below 45.0 μg/m^3).

The forecasting of PM2.5 concentration is significantly impacted by topographical features which affect both the transport of PM2.5 and the stability of the atmosphere. We compared the forecasting accuracy with and without consideration of terrain for 12, 18, and 24-hr time frames (Table 1). When only historical PM2.5 and meteorological features are considered in model training and forecasting processes, forecasting accuracy decreases rapidly over time. Specifically, the *R* decreases from 0.86 at 6 hr to 0.19 at 24 hr, while the RMSE increases from 16.74 to 100.49 μg/m^3. In contrast, when the terrain was incorporated into the forecasting model, the *R* remains above 0.85 within 24 hr and the RMSE slowly increases from 26.72 μg/m^3 at 6 hr to 30.57 μg/m^3 at 24 hr.

Table 1
Comparison of PM2.5 Forecasting Using Different Combinations of Input Features

	6 hr		12 hr		24 hr	
Input features \ lead time	*R*	RMSE	*R*	RMSE	*R*	RMSE
hPM2.5	0.85	22.78	0.31	28.97	–	–
hPM2.5 + meteo-features	0.86	16.74	0.66	21.58	0.19	100.49
hPM2.5 + meteo & topo-features	0.90	26.72	0.85	29.97	0.88	30.57

Table 2
Comparison of DA Accuracy and Execution Time Using EnKF With Different Ensemble Size

Ensemble size	Radius	RMSE	*R*	Time
50	100	13.57	0.79	23.92 s
100	500	10.54	0.85	46.44 s
256	1,000	9.91	0.87	4 min
1,000	5,000	7.49	0.93	8 min

4.4. The Effect of Ensemble Size on the Accuracy of DA Result

The EnKF algorithm is susceptible to spurious correlations which stem from utilizing a limited ensemble size. This can cause inaccuracies in representing the actual probability density function of the system being studied. Consequently, the EnKF may generate inaccurate estimations that do not reflect the genuine state of the system, which can result in incorrect results.

To mitigate this problem, it is suggested to increase the ensemble size since the use of the ConvLSTM model for forecast can largely improve computational efficiency. We have compared the ConvLSTM-EnKF and NAQPMS-EnKF methods in terms of the CPU time required for DA. The ConvLSTM-EnKF method can reduce the CPU time by three orders of magnitude compared to the NAQPMS-EnKF method, a typical physical model-based method. More information on computational efficiency can be found in Appendix in Supporting Information S1 Text S1 in Supporting Information S1. Data-driven methods enable larger ensemble sizes with restricted computational resources, which can enhance the accuracy of DA.

We examined the efficacy of DA using the EnKF with varying ensemble sizes of 50, 100, 256, and 1,000, as illustrated in Table 2. We applied localization in EnKF to avoid spurious correlations and tuned the radius for DA studies with different ensemble sizes. In our study, the localization radius used is increased with ensemble size (Kirchgessner et al., 2014). The accuracy of DA improves as the ensemble size increases. With only 50 ensemble members, the RMSE and R between DA results and observations are 13.57 $\mu g/m^3$ and 0.79, respectively. Increasing the ensemble size to 100 led to a notable reduction in RMSE to 10.54 $\mu g/m^3$ and an increase in R to 0.85. Additionally, a single DA cycle only requires 46.44 s. Notably, the proposed ConvLSTM-EnKF method can accommodate even larger ensemble sizes. By employing 1,000 ensemble members, the RMSE was further reduced to 7.49 $\mu g/m^3$ and R increased to 0.93 while requiring only 8 min for a single DA cycle.

5. Conclusions

In this study, we have developed a ConvLSTM-EnKF hybrid model for accurate and efficient online real-time forecasting at high spatial and temporal resolutions. This model combines a purely data-driven ConvLSTM model with a DA method (EnKF) to produce reliable operational spatiotemporal forecasting in complex high-dimensional systems. The ConvLSTM model is advantageous as it simultaneously captures spatial and temporal correlations within the data, while EnKF is used to correct forecasting errors by assimilating real-time observations and tackling cumulative errors.

The capability of the proposed method is demonstrated in PM2.5 forecasting across the entire China. This is a challenging task due to the complex topographical and meteorological conditions in China, the need for high-resolution forecasting over a large study area with hourly time intervals, and the scarcity of observations. The key findings and conclusions of the study are presented below:

1. *Operational forecasting*: The ConvLSTM-EnKF method has a good performance in operational forecasting, delivering consistent and stable forecasts for up to 1 month. Our findings demonstrate that the ConvLSTM-EnKF forecasts are consistent with observations in March, June, September, and December 2018.
2. *Spatial and temporal correlations*: Our results also indicate that the trained ConvLSTM model can capture both the spatial and temporal dependences of PM2.5 concentrations, with the majority of China having absolute prediction errors smaller than 10 $\mu g/m^3$. The largest prediction error with an absolute value of 40 $\mu g/m^3$ occurred in the Jing-Jin-Ji region which has a high population density and developed industries.
3. *Online DA at a high spatial resolution with sparse observations*: The ConvLSTM-EnKF enables online DA at a high spatial resolution and produces effective forecasting and DA results, while also allowing for the utilization of larger ensemble sizes, which improves the accuracy of DA results.
4. *The efficiency of forecasting and DA*: The CPU time required for the proposed ConvLSTM-EnKF forecasting is reduced by three orders of magnitude compared to the conventional NAQPMS-EnKF method.

Overall, the hybrid purely data-driven ConvLSTM-EnKF provides efficient and accurate hourly PM2.5 forecasting in China. This paves the way toward operational real-time prediction and management. Future work will focus on explainable AI as well as physical-informed machine learning modeling.

Data Availability Statement

The ConvLSTM is a public model and can be found at TensorFlow Developers (2023). The CAQRA data set is available from Tang et al. (2021). The monitoring air pollution data used in this study were accessed through the Ministry of Ecology and Environment of the People's Republic of China (http://www.mee.gov.cn/).

Acknowledgments
The authors are grateful to the sponsors of the Resource Geophysics Academy, Imperial College London for supporting this research. This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) (MAGIC (EP/N010221/1), INHALE (EP/T003189/1) and PREMIERE (EP/X029093/1, EP/T000414/1) in the UK. The authors acknowledge the reviewers, editors, and Dr. Xiaofei Wu for their in-depth perspicacious comments that contributed to improving the presentation of this paper.

References

Alléon, A., Jauvion, G., Quennehen, B., & Lissmyr, D. (2020). PlumeNet: Large-scale air quality forecasting using a convolutional LSTM network. ArXiv Preprint ArXiv:2006.09204. Retrieved from http://arxiv.org/abs/2006.09204

An, Z., Huang, R. J., Zhang, R., Tie, X., Li, G., Cao, J., et al. (2019). Severe haze in northern China: A synergy of anthropogenic emissions and atmospheric processes. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(18), 8657–8666. https://doi.org/10.1073/pnas.1900125116

Bocquet, M., Elbern, H., Eskes, H., Hirtl, M., Aabkar, R., Carmichael, G. R., et al. (2015). Data assimilation in atmospheric chemistry models: Current status and future prospects for coupled chemistry meteorology models. *Atmospheric Chemistry and Physics*, *15*(10), 5325–5358. https://doi.org/10.5194/acp-15-5325-2015

Brunner, M. I., Gilleland, E., Wood, A., Swain, D. L., & Clark, M. (2020). Spatial dependence of floods shaped by spatiotemporal variations in meteorological and land-surface processes. *Geophysical Research Letters*, *47*(13), 1–13. https://doi.org/10.1029/2020GL088000

Chai, S., Xu, Z., Jia, Y., & Wong, W. K. (2020). A robust spatiotemporal forecasting framework for photovoltaic generation. *IEEE Transactions on Smart Grid*, *11*(6), 5370–5382. https://doi.org/10.1109/TSG.2020.3006085

Chattopadhyay, A., Mustafa, M., Hassanzadeh, P., Bach, E., & Kashinath, K. (2022). Towards physics-inspired data-driven weather forecasting: Integrating data assimilation with a deep spatial-transformer-based U-NET in a case study with ERA5. *Geoscientific Model Development*, *15*(5), 2221–2237. https://doi.org/10.5194/gmd-15-2221-2022

Cheng, M., Fang, F., Navon, I. M., Zheng, J., Tang, X., Zhu, J., & Pain, C. (2022). Spatio-temporal hourly and daily ozone forecasting in China using a hybrid machine learning model: Autoencoder and generative adversarial networks. *Journal of Advances in Modeling Earth Systems*, *14*(3), 1–26. https://doi.org/10.1029/2021MS002806

Dash, U. K., Park, S. Y., Song, C. H., Yu, J., Yumimoto, K., & Uno, I. (2023). Performance comparisons of the three data assimilation methods for improved predictability of PM2·5: Ensemble Kalman filter, ensemble square root filter, and three-dimensional variational methods. *Environmental Pollution*, *322*, 121099. https://doi.org/10.1016/j.envpol.2023.121099

Elbern, H., & Schmidt, H. (2001). Ozone episode analysis by four-dimensional variational chemistry data assimilation. *Journal of Geophysical Research*, *106*(D4), 3569–3590. https://doi.org/10.1029/2000JD900448

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211. https://doi.org/10.1016/0364-0213(90)90002-E

Evensen, G. (1994). Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Research*, *99*(C5), 10143–10162. https://doi.org/10.1029/94JC00572

Evensen, G., Vossepoel, F. C., & van Leeuwen, P. J. (2022). *Data assimilation fundamentals: A unified formulation of the state and parameter estimation problem*. Springer. Retrieved from https://link.springer.com/10.1007/978-3-030-96709-3

Ghil, M., & Malanotte-Rizzoli, P. (1991). Data assimilation in meteorology and oceanography. In *Advances in Geophysics* (Vol. 33, pp. 141–266). Elsevier. https://doi.org/10.1016/S0065-2687(08)60442-2

Hakim, G. J., Emile-Geay, J., Steig, E. J., Noone, D., Anderson, D. M., Tardif, R., et al. (2016). The last millennium climate reanalysis project: Framework and first results. *Journal of Geophysical Research*, *121*(12), 6745–6764. https://doi.org/10.1002/2016JD024751

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hong, C., Zhang, Q., Zhang, Y., Davis, S. J., Tong, D., Zheng, Y., et al. (2019). Impacts of climate change on future air quality and human health in China. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(35), 17193–17200. https://doi.org/10.1073/pnas.1812881116

Houtekamer, P. L., He, B., & Mitchell, H. L. (2014). Parallel implementation of an ensemble Kalman filter. *Monthly Weather Review*, *142*(3), 1163–1182. https://doi.org/10.1175/MWR-D-13-00011.1

Houtekamer, P. L., & Zhang, F. (2016). Review of the ensemble Kalman filter for atmospheric data assimilation. *Monthly Weather Review*, *144*(12), 4489–4532. https://doi.org/10.1175/MWR-D-15-0440.1

Karevan, Z., & Suykens, J. A. K. (2020). Transductive LSTM for time-series prediction: An application to weather forecasting. *Neural Networks*, *125*, 1–9. https://doi.org/10.1016/j.neunet.2019.12.030

Kim, S., Hong, S., Joh, M., & Song, S. (2017). In *DeepRain: ConvLSTM network for precipitation prediction using multichannel radar data* (Vol. 3–6). Retrieved from http://arxiv.org/abs/1711.02316

Kirchgessner, P., Nerger, L., & Bunse-Gerstner, A. (2014). On the choice of an optimal localization radius in ensemble Kalman filter methods. *Monthly Weather Review*, *142*(6), 2165–2175. https://doi.org/10.1175/MWR-D-13-00246.1

Kochkov, D., Smith, J. A., Alieva, A., Wang, Q., Brenner, M. P., & Hoyer, S. (2021). Machine learning–accelerated computational fluid dynamics. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(21). https://doi.org/10.1073/pnas.2101784118

Kong, L., Tang, X., Zhu, J., Wang, Z., Li, J., Wu, H., et al. (2021). A 6-year-long (2013–2018) high-resolution air quality reanalysis dataset in China based on the assimilation of surface observations from CNEMC. *Earth System Science Data*, *13*(2), 529–570. https://doi.org/10.5194/essd-13-529-2021

Leutbecher, M. (2019). Ensemble size: How suboptimal is less than infinity? *Quarterly Journal of the Royal Meteorological Society*, *145*(S1), 107–128. https://doi.org/10.1002/qj.3387

Li, X., Peng, L., Yao, X., Cui, S., Hu, Y., You, C., & Chi, T. (2017). Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation. *Environmental Pollution*, *231*, 997–1004. https://doi.org/10.1016/j.envpol.2017.08.114

Liang, F., Xiao, Q., Huang, K., Yang, X., Liu, F., Li, J., et al. (2020). The 17-y spatiotemporal trend of PM 2.5 and its mortality burden in China. *Proceedings of the National Academy of Sciences of the United States ofAmerica*, *117*(41), 25601–25608. https://doi.org/10.1073/pnas.1919641117

Liu, Y., Qin, H., Zhang, Z., Pei, S., Wang, C., Yu, X., et al. (2019). Ensemble spatiotemporal forecasting of solar irradiation using variational Bayesian convolutional gate recurrent unit network. *Applied Energy*, *253*(July), 113596. https://doi.org/10.1016/j.apenergy.2019.113596

Lorenc, A. C. (2003). The potential of the ensemble Kalman filter for NWP—A comparison with 4D-Var. *Quarterly Journal of the Royal Meteorological Society*, *129*(595), 3183–3203. https://doi.org/10.1256/qj.02.132

Lorenz, E. N. (1963). Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, *20*(2), 130–141. https://doi.org/10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2

Lorenz, E. N. (1972). Predictability: Does the flap of a butterfly's wings in Brazil set off a tornado in Texas? In *American association for the advancement of science* (pp. 11–13). American Association for the Advancement of Science.

Niu, M., Zhang, Y., & Ren, Z. (2023). Deep learning-based PM2.5 long time-series prediction by fusing multisource data—A case study of Beijing. *Atmosphere*, *14*(2), 340. https://doi.org/10.3390/atmos14020340

Park, S. K., & Xu, L. (2022). In S. K. Park & L. Xu (Eds.), *Data assimilation for atmospheric, oceanic and hydrologic applications* (Vol. IV). Springer International Publishing. https://doi.org/10.1007/978-3-030-77722-7

Pawar, S., & San, O. (2022). Equation-free surrogate modeling of geophysical flows at the intersection of machine learning and data assimilation. *Journal of Advances in Modeling Earth Systems*, *14*(11), 1–20. https://doi.org/10.1029/2022MS003170

Peng, Z., Liu, Z., Chen, D., & Ban, J. (2017). Improving PM2.5 forecast over China by the joint adjustment of initial conditions and source emissions with an ensemble Kalman filter. *Atmospheric Chemistry and Physics*, *17*(7), 4837–4855. https://doi.org/10.5194/acp-17-4837-2017

Penny, S. G., Smith, T. A., Chen, T. C., Platt, J. A., Lin, H. Y., Goodliff, M., & Abarbanel, H. D. I. (2022). Integrating recurrent neural networks with data assimilation for scalable data-driven state estimation. *Journal of Advances in Modeling Earth Systems*, *14*(3), 1–25. https://doi.org/10.1029/2021MS002843

Qi, Y., Li, Q., Karimian, H., & Liu, D. (2019). A hybrid model for spatiotemporal forecasting of PM 2.5 based on graph convolutional neural network and long short-term memory. *Science of the Total Environment*, *664*, 1–10. https://doi.org/10.1016/j.scitotenv.2019.01.333

Rasp, S., & Thuerey, N. (2021). Data-driven medium-range weather prediction with a resnet pretrained on climate simulations: A new model for WeatherBench. *Journal of Advances in Modeling Earth Systems*, *13*(2). https://doi.org/10.1029/2020MS002405

Ravuri, S., Lenc, K., Willson, M., Kangin, D., Lam, R., Mirowski, P., et al. (2021). Skilful precipitation nowcasting using deep generative models of radar. *Nature*, *597*(7878), 672–677. https://doi.org/10.1038/s41586-021-03854-z

Saha, S., Moorthi, S., Pan, H. L., Wu, X., Wang, J., Nadiga, S., et al. (2010). The NCEP climate forecast system reanalysis. *Bulletin of the American Meteorological Society*, *91*(8), 1015–1057. https://doi.org/10.1175/2010BAMS3001.1

Sandu, A., & Chai, T. (2011). Chemical data assimilation-An overview. *Atmosphere*, *2*(3), 426–463. https://doi.org/10.3390/atmos2030426

Schemm, S., Grund, D., Knutti, R., Wernli, H., Ackermann, M., & Evensen, G. (2023). Learning from weather and climate science to prepare for a future pandemic. *Proceedings of the National Academy of Sciences of the United States of America*, *120*(4), 17–20. https://doi.org/10.1073/pnas.2209091120

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W., & Woo, W. (2015). Convolutional LSTM 730 network: A machine learning approach for precipitation nowcasting. *Advances in Neural 731 Information Processing Systems*, *28*. Retrieved from http://arxiv.org/abs/1506.04214

Shi, X., & Yeung, D.-Y. (2018). Machine learning for spatiotemporal sequence forecasting: A survey. Retrieved from http://arxiv.org/abs/1808.06865

Skamarock, W. C., Klemp, J. B., Dudhia, J., Gill, D. O., Barker, D., Duda, M. G., et al. (2008). *A description of the advanced Research WRF version 3*. University Corporation for Atmospheric Research. https://doi.org/10.5065/D68S4MVH

Song, Y., & Shao, M. (2023). Impacts of complex terrain features on local wind field and PM2.5 concentration. *Atmosphere*, *14*(5), 761. https://doi.org/10.3390/atmos14050761

Tang, X., Kong, L., Zhu, J., Wang, Z., Li, J., Wu, Q., et al. (2021). A high-resolution air quality reanalysis dataset over China (CAQRA) (version 3) [Dataset]. Science Data Bank. https://doi.org/10.11922/sciencedb.00053

Teng, M., Li, S., Xing, J., Fan, C., Yang, J., Wang, S., et al. (2023). 72-Hour real-time forecasting of ambient Pm2.5 by hybrid graph deep neural network with aggregated neighborhood spatiotemporal information. *The Social Science Research Network*. https://doi.org/10.2139/ssrn.4355589

TensorFlow Developers. (2023). TensorFlow (version 2.14.0) [Software]. Zenodo. https://doi.org/10.5281/zenodo.8381573

Weyn, J. A., Durran, D. R., & Caruana, R. (2019). Can machines learn to predict weather? Using deep learning to predict gridded 500-hPa geopotential height from historical weather data. *Journal of Advances in Modeling Earth Systems*, *11*(8), 2680–2693. https://doi.org/10.1029/2019MS001705

Xu, L., Chen, N., Chen, Z., Zhang, C., & Yu, H. (2021). Spatiotemporal forecasting in earth system science: Methods, uncertainties, predictability and future directions. *Earth-Science Reviews*, *222*, 103828. https://doi.org/10.1016/j.earscirev.2021.103828

Xu, Y., Ho, H. C., Wong, M. S., Deng, C., Shi, Y., Chan, T. C., & Knudby, A. (2018). Evaluation of machine learning techniques with multiple remote sensing datasets in estimating monthly concentrations of ground-level PM2.5. *Environmental Pollution*, *242*, 1417–1426. https://doi.org/10.1016/j.envpol.2018.08.029

Yan, R., Liao, J., Yang, J., Sun, W., Nong, M., & Li, F. (2021). Multi-hour and multi-site air quality index forecasting in Beijing using CNN, LSTM, CNN-LSTM, and spatiotemporal clustering. *Expert Systems with Applications*, *169*, 114513. https://doi.org/10.1016/j.eswa.2020.114513

Zhang, Q., Zheng, Y., Tong, D., Shao, M., Wang, S., Zhang, Y., et al. (2019). Drivers of improved PM2.5 air quality in China from 2013 to 2017. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(49), 24463–24469. https://doi.org/10.1073/pnas.1907956116

Zhu, F., Emile-Geay, J., McKay, N. P., Hakim, G. J., Khider, D., Ault, T. R., et al. (2019). Climate models can correctly simulate the continuum of global-average temperature variability. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(18), 8728–8733. https://doi.org/10.1073/pnas.1809959116

References From the Supporting Information

Ben Taieb, S., Sorjamaa, A., & Bontempi, G. (2010). Multiple-output modeling for multi-step-ahead time series forecasting. *Neurocomputing*, *73*(10–12), 1950–1957. https://doi.org/10.1016/j.neucom.2009.11.030

Bontempi, G., & Ben Taieb, S. (2011). Conditionally dependent strategies for multiple-step-ahead prediction in local learning. *International Journal of Forecasting*, *27*(3), 689–699. https://doi.org/10.1016/j.ijforecast.2010.09.004

Cheng, M., Fang, F., Pain, C. C., & Navon, I. M. (2020). Data-driven modelling of nonlinear spatio-temporal fluid flows using a deep convolutional generative adversarial network. *Computer Methods in Applied Mechanics and Engineering*, *365*, 113000. https://doi.org/10.1016/j.cma.2020.113000

Evensen, G. (2003). The ensemble Kalman filter: Theoretical formulation and practical implementation. *Ocean Dynamics*, *53*(4), 343–367. https://doi.org/10.1007/s10236-003-0036-9

Graves, A. (2013). In *Generating sequences with recurrent neural networks* (Vol. 1–43). Retrieved from http://arxiv.org/abs/1308.0850

Lopez-Restrepo, S., Yarce, A., Pinel, N., Quintero, O. L., Segers, A., & Heemink, A. W. (2020). Forecasting PM10 and PM2.5 in the Aburrá Valley (Medellín, Colombia) via EnKF based data assimilation. *Atmospheric Environment*, *232*, 117507. https://doi.org/10.1016/j.atmosenv.2020.117507