Data-Driven Methods for Weather Forecast

Elias D. Nino-Ruiz[0000−0001−7784−8163] and
Felipe J. Acevedo García[0000−0002−2148−7167]

Applied Math and Computer Science Lab, Computer Science Department,
Universidad del Norte, Colombia {enino,fjacevedo}@uninorte.edu.co
https://sites.google.com/vt.edu/eliasn/home?authuser=1

Abstract. In this paper, we propose efficient and practical data-driven methods for weather forecasts. We exploit the information brought by historical weather datasets to build machine-learning-based models. These models are employed to produce numerical forecasts, which can be improved by injecting additional data via data assimilation. Our approaches' general idea is as follows: given a set of time snapshots of some dynamical system, we group the data by time across multiple days. These groups are employed to build first-order Markovian models that reproduce dynamics from time to time. Our numerical models' precision can be improved via sequential data assimilation. Experimental tests are performed by using the National-Centers-for-Environmental-Prediction Department-of-Energy Reanalysis II dataset. The results reveal that numerical forecasts can be obtained within reasonable error magnitudes in the L_2 norm sense, and even more, observations can improve forecasts by order of magnitudes, in some cases.

Keywords: Data Assimilation · Markovian Model · Machine Learning.

1 Introduction

Numerical weather forecasts are of extreme importance in different aspects of life, particularly in scenarios wherein human lives can be compromised (i.e., forecasts of storms, floods, hurricanes, and tsunamis) [10]. Numerical models are commonly employed to mimic the behavior of actual system dynamics, for instance, the ocean and/or the atmosphere [6, 2, 9]. Since numerical models are computationally demanding, high-performance-computing is a must to produce forecasts within reasonable computational times, especially for high-resolution grids. On the other hand, we can find data-driven models that can exploit decades of meteorological information to represent the future as some potential combination of the past. For instance, the National-Centers-for-Environmental-Prediction Department-of-Energy (NCEP-DOE) Reanalysis II [5] is a data set that holds meteorological information since 1979 onto global grids at varying resolutions. It is possible to use these sources of information to come up with statistical forecasts. We think there is an opportunity to compute cheap models that produce forecasts with low computational effort, and even more, we can improve such forecasts by injecting real-time data via sequential data assimilation.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77970-2_25

This paper is organized as follows: in Section 2 we discuss topics related to sequential data assimilation and machine learning models, Section 3 presents a sequential data assimilation method via machine learning models wherein numerical models are replaced by statistical ones, in Section 4 experiments are carried out by employing the NCEP-DOE Reanalysis II data set, and the conclusions of this research are stated in Section 5.

2 Preliminaries

In this section, we discuss some topics related to Machine Learning and Sequential Data Assimilation methods. These are necessary for the understanding of our proposed methods.

2.1 Machine Learning Models

In the context of Machine Learning (ML), parametric models can be seen as structures for the solution of the inverse problem [11]:

$$y = f\left(\boldsymbol{\beta}, \mathbf{z}\right) + \epsilon\,, \tag{1}$$

where y is an observation, $\boldsymbol{\beta} \in \mathbb{R}^{p\times 1}$ is a vector holding the parameters, p is the number of parameters, variables are stored in vector $\mathbf{z} \in \mathbb{R}^{v\times 1}$, v is the number of variables, $f : \mathbb{R}^{p\times 1} \times \mathbb{R}^{v\times 1} \to 1$, and ϵ can be described by some probability density function. The simplest model in which one can think is a linear one of the form:

$$y = \sum_{u=1}^{v} \beta_u \cdot z_u + \epsilon, \text{ with } \epsilon \sim \mathcal{N}\left(0, \sigma^2\right)\,, \tag{2}$$

where $v = p$, β_u and z_u denote the u-th component of vectors $\boldsymbol{\beta}$ and $\mathbf{z}$, respectively. Since Gaussian assumptions can be easily broken on residuals in (9), local linear models can be built to preserve Gaussian shapes (i.e., by considering, local modes of error distributions). There are many manners to do this, the simpler, to weight each sample $\mathbf{x}_j \in \mathbb{R}^{v\times 1}$, for $1 \leq j \leq m$, with regard to its distance to the observations $\mathbf{y} \in \mathbb{R}^{m\times 1}$,

$$\mathbf{y} = \sum_{j=1}^{K<m} \beta_j \cdot \alpha_j(\mathbf{x}_j, \mathbf{y}) \cdot \mathbf{x}_j + \boldsymbol{\epsilon}, \text{ with } \boldsymbol{\epsilon} \sim \mathcal{N}\left(0, \sigma^2 \cdot \mathbf{I}\right)\,, \tag{3a}$$

where m is the number of samples (observations), K is the number of closest points to $\mathbf{y}$ onto the hyperplane formed by the samples $\mathbf{x}_j$, $\boldsymbol{\epsilon}$ is a vector holding the residuals. Likewise, $\alpha_j(\mathbf{x}_s, \mathbf{y})$ is a weight/distance function, common choices are the Uniform distance

$$\alpha_j(\mathbf{x}_j, \mathbf{y})_\infty = \|\mathbf{x}_j - \mathbf{y}\|_\infty^{-1}\,, \tag{3b}$$

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77970-2_25

or the reciprocal of the Euclidean distance (L_2-norm):

$$\alpha_j(\mathbf{x}_j, \mathbf{y})_2 = \|\mathbf{x}_j - \mathbf{y}\|_2^{-1} \,. \tag{3c}$$

This kind of strategy is well-known in the statistical context as the K-Nearest-Neighbors (KNN) regression [12].

2.2 Sequential Data Assimilation

The ensemble Kalman filter (EnKF) is a well-established sequential Monte Carlo method for parameter and state estimation in highly non-linear models [3, 4]. The EnKF describes the error statistics via an ensemble of model realizations:

$$\mathbf{X}^b = \left[\mathbf{x}^{b[1]}, \mathbf{x}^{b[2]}, \ldots, \mathbf{x}^{b[N]}\right] \in \mathbb{R}^{n \times N} \,, \tag{4}$$

where $\mathbf{x}^{b[e]}$, for $1 \leq e \leq N$, is the e-th ensemble member, N is the ensemble size, and n denotes the model dimension. The background ensemble (4) can be employed to estimate the moments of the background error distribution, this is, the background state $\mathbf{x}^b \in \mathbb{R}^{n \times 1}$

$$\mathbf{x}^b \approx \overline{\mathbf{x}}^b = \frac{1}{N} \cdot \sum_{e=1}^{N} \mathbf{x}^{b[e]} \in \mathbb{R}^{n \times 1} \,, \tag{5a}$$

and the background error covariance matrix $\mathbf{B} \in \mathbb{R}^{n \times n}$

$$\mathbf{B} \approx \mathbf{P}^b = \frac{1}{N-1} \cdot \boldsymbol{\Delta}\mathbf{X}^b \cdot \left[\boldsymbol{\Delta}\mathbf{X}^b\right]^T \in \mathbb{R}^{n \times n} \,, \tag{5b}$$

where $\overline{\mathbf{x}}^b$ is the ensemble mean, and $\mathbf{P}^b$ is the ensemble covariance. Likewise, $\boldsymbol{\Delta}\mathbf{X}^b \in \mathbb{R}^{n \times N}$ stands for the matrix of member deviations $\boldsymbol{\Delta}\mathbf{X}^b = \mathbf{X}^b - \overline{\mathbf{x}}^b \cdot \mathbf{1}_N^T$ wherein $\mathbf{1}_N$ is an N-dimensional vector whose components are all ones. Observations are related to model states via the linear observation operator $\mathbf{H} \in \mathbb{R}^{m \times n}$

$$\mathbf{y} = \mathbf{H} \cdot \mathbf{x} + \boldsymbol{\varepsilon} \in \mathbb{R}^{m \times 1} \,,$$

where $\mathbf{H}$ maps model states onto observation spaces, m is the number of observations, the white noise vector reads $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$, and $\mathbf{R} \in \mathbb{R}^{m \times m}$ is the data error covariance matrix. By using Bayes' theorem, we can find the state that maximizes the posterior probability given an observation $\mathbf{y}$ as follows:

$$\mathbf{x}^a = \mathbf{x}^b + \mathbf{P}^b \cdot \mathbf{H}^T \cdot \left[\mathbf{R} + \mathbf{H} \cdot \mathbf{P}^b \cdot \mathbf{H}^T\right] \cdot \left[\mathbf{y} - \mathbf{H} \cdot \mathbf{x}^b\right] \in \mathbb{R}^{n \times 1} \,, \tag{6}$$

Since model realizations come at high computational costs, ensemble sizes are constrained by the hundreds while model resolutions range in the order of millions. This tigers spurious correlations between errors in distant model components. To mitigate this, better estimations of $\mathbf{B}$ are sought. For instance, sparse precision covariances of the form can be computed as follows:

$$\widehat{\mathbf{B}}^{-1} = \widehat{\mathbf{L}}^T \cdot \widehat{\mathbf{D}}^{-1} \cdot \widehat{\mathbf{L}} \in \mathbb{R}^{n \times n} \tag{7}$$

where $\widehat{\mathbf{L}} \in \mathbb{R}^{n \times n}$ is a sparse lower triangular matrix, and $\widehat{\mathbf{D}}^{-1}$ is a diagonal matrix [1]. By replacing (7) in (6) the EnKF via a modified Cholesky decomposition [7, 8] can be obtained.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77970-2_25

3 Proposed Method

Consider time snapshots of some physical variables onto a numerical mesh grid (i.e., the NCEP-DOE Reanalysis II data set). We group the snapshots (data) by time across different S days:

$$\mathbf{X}_\ell = \left[\mathbf{x}_\ell^{[1]}, \mathbf{x}_\ell^{[2]}, \ldots, \mathbf{x}_\ell^{[S]}\right] \in \mathbb{R}^{n \times S}\,, \tag{8}$$

where $\mathbf{X}_\ell$ is the ensemble holding all snapshots at time t_ℓ across different days, for $1 \leq \ell \leq L$, L is the number of snapshots in a single day, $\mathbf{x}_\ell^{[s]} \in \mathbb{R}^{n \times 1}$ is the snapshot of day s, for $1 \leq s \leq S$, at time t_ℓ, and S is the number of days (number of snapshots for time t_ℓ across different days). We consider evenly-spaced time snapshots, and we assume that the same number of snapshots are available for all ensembles $\mathbf{X}_\ell$. We then consider to fit models of the form:

$$\mathbf{X}_\ell = \mathbf{M}_{\ell,\ell-1} \cdot \mathbf{X}_{\ell-1} + \mathbf{E}_\ell\,, \tag{9}$$

where $\mathbf{E}_\ell \in \mathbb{R}^{n \times S}$ holds the residuals, and $\mathbf{M}_{\ell,\ell-1} \in \mathbb{R}^{n \times n}$ is a data-driven model which partially captures the evolution of system dynamics from time $t_{\ell-1}$ to t_ℓ. We then consider cost functions of the form:

$$\mathcal{J}\left(\mathbf{M}_{\ell,\ell-1}\right) = \frac{1}{2} \cdot \|\mathbf{X}_\ell - \mathbf{M}_{\ell,\ell-1} \cdot \mathbf{X}_{\ell-1}\|_2^2\,, \tag{10}$$

to estimate linear operators $\mathbf{M}_{\ell,\ell-1}$ which transport dynamics from time $t_{\ell-1}$ to t_ℓ, the optimization problem to solve reads:

$$\mathbf{M}_{\ell,\ell-1}^* = \arg\min_{\mathbf{M}_{\ell,\ell-1}} \mathcal{J}\left(\mathbf{M}_{\ell,\ell-1}\right)\,. \tag{11}$$

It can be easily shown that the gradient of (10) reads

$$\nabla_{\mathbf{M}_{\ell,\ell-1}}\left(\mathcal{J}\left(\mathbf{M}_{\ell,\ell-1}\right)\right) = \mathbf{X}_\ell \cdot \left[\mathbf{X}_{\ell-1}\right]^T - \mathbf{M}_{\ell,\ell-1} \cdot \mathbf{X}_{\ell-1} \cdot \left[\mathbf{X}_{\ell-1}\right]^T\,, \tag{12}$$

from which the solution of (11) reads:

$$\mathbf{M}_{\ell,\ell-1}^* = \mathbf{X}_\ell \cdot \left[\mathbf{X}_{\ell-1}\right]^T \cdot \left[\mathbf{X}_{\ell-1} \cdot \left[\mathbf{X}_{\ell-1}\right]^T\right]^{-1}\,. \tag{13}$$

We now have a piecewise linear model $\left\{\mathbf{M}_{\ell,\ell-1}^*\right\}_{\ell=1}^{L}$ which mimics the behavior of the dynamical system in a single day. Note that, these set of models can be seen as first order Markovian models wherein all information needed to propagate dynamics from time $t_{\ell-1}$ to t_ℓ is condensed into $\mathbf{M}_{\ell-1,\ell}$.

3.1 Sequential Data Assimilation

Since linear models of the form (13) can partially capture the actual dynamics, we can improve their accuracies by using sequential data assimilation. For each

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77970-2_25

time t_ℓ, we can build the precision matrix $\widehat{\mathbf{B}}_\ell^{-1}$ via the ensemble (8) and the modified Cholesky decomposition:

$$\widehat{\mathbf{B}}_\ell^{-1} = \widehat{\mathbf{L}}_\ell^T \cdot \widehat{\mathbf{D}}_\ell^{-1} \cdot \widehat{\mathbf{L}}_\ell \in \mathbb{R}^{n \times n}\,, \tag{14}$$

where the factors $\widehat{\mathbf{L}}_\ell$, and $\widehat{\mathbf{D}}_\ell^{-1}$ are computed as in Section 2. We model the uncertinties of any state at time t_ℓ by (14). Consider the analysis state $\mathbf{x}_{\ell-1}^a$, it is clear that:

$$\mathbf{x}_\ell^b = \mathbf{M}_{\ell,\ell-1}^* \cdot \mathbf{x}_{\ell-1}^a = \mathbf{X}_\ell \cdot [\mathbf{X}_{\ell-1}]^T \cdot \left[\mathbf{X}_{\ell-1} \cdot [\mathbf{X}_{\ell-1}]^T\right]^{-1} \cdot \mathbf{x}_{\ell-1}^a\,. \tag{15}$$

Consider the observation $\mathbf{y}_\ell \in \mathbb{R}^{m\times 1}$, at time t_ℓ , the analysis state can be computed as follows:

$$\begin{aligned}\mathbf{x}_\ell^a = &\left[\widehat{\mathbf{B}}_\ell^{-1} + \mathbf{H}_\ell^T \cdot \mathbf{R}_\ell^{-1} \cdot \mathbf{H}_\ell\right]^{-1} \\ &\cdot \left[\widehat{\mathbf{B}}_\ell^{-1} \cdot \mathbf{X}_\ell \cdot [\mathbf{X}_{\ell-1}]^T \cdot \left[\mathbf{X}_{\ell-1} \cdot [\mathbf{X}_{\ell-1}]^T\right]^{-1} \cdot \mathbf{x}_{\ell-1}^a + \mathbf{H}_\ell^T \cdot \mathbf{R}_\ell^{-1} \cdot \mathbf{y}_\ell\right]\end{aligned} \tag{16}$$

The analysis state (16) is propagated until new observations are available.

3.2 Building Local Linear Models

In practice, model dynamics can be highly non-linear and therefore, Gaussian assumptions on residuals in (9) can be easily broken. To mitigate this, consider the ensemble $\mathbf{X}(\mathbf{x}_{\ell-1}^a, K) \in \mathbb{R}^{n\times K}$ formed by the K nearest states to $\mathbf{x}_{\ell-1}^a$ from the ensemble of snapshots $\mathbf{X}_{\ell-1}$ at time $t_{\ell-1}$: the ensemble $\mathbf{X}(\mathbf{x}_{\ell-1}^a, K)$ can be exploited to build local linear models

$$\mathbf{x}_\ell^b = \mathbf{M}\left(\mathbf{x}_{\ell-1}^a,\, K\right)_{\ell,\,\ell-1}^* \cdot \mathbf{x}_{\ell-1}^a\,, \tag{17a}$$

following a similar reasoning to that of (3). This is, by considering the reciprocal of Uniform or Euclidean distances. Besides, K can be employed as well to fit local Gaussian models for prior errors by using the forecast state (15), and its K nearest states (neighbors) in (8), $\mathbf{X}(\mathbf{x}_\ell^b, K) \in \mathbb{R}^{n\times K}$; from here we can obtain local approximations of precision matrices of the form (14):

$$\widehat{\mathbf{B}}^{-1}\left(\mathbf{x}_\ell^b,\, K\right)_\ell = \widehat{\mathbf{L}}\left(\mathbf{x}_\ell^b,\, K\right)_\ell^T \cdot \widehat{\mathbf{D}}\left(\mathbf{x}_\ell^b,\, K\right)_\ell^{-1} \cdot \widehat{\mathbf{L}}\left(\mathbf{x}_\ell^b,\, K\right)_\ell \in \mathbb{R}^{n\times n}\,, \tag{17b}$$

where $\widehat{\mathbf{L}}\left(\mathbf{x}_\ell^b,\, K\right)_\ell^T \in \mathbb{R}^{n\times n}$, and $\widehat{\mathbf{D}}\left(\mathbf{x}_\ell^b,\, K\right)_\ell^{-1} \in \mathbb{R}^{n\times n}$ are computed similar to those in (14). We decide to make use of the precision matrix $\widehat{\mathbf{B}}^{-1}$ in our formulation since it allows us to exploit the use of computational resources (the resulting estimator is sparse).

The background state (17a) and the precision matrix (17b) can be easily incorporated into the data assimilation framework (16), this is, we can produce

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77970-2_25

forecasts via (17a), and analysis states:

$$\mathbf{x}_{\ell}^{a}(K) = \left[\widehat{\mathbf{B}}^{-1}\left(\mathbf{x}_{\ell}^{b},\, K\right)_{\ell} + \mathbf{H}_{\ell}^{T}\cdot\mathbf{R}_{\ell}^{-1}\cdot\mathbf{H}_{\ell}\right]^{-1} \cdot \left[\widehat{\mathbf{B}}^{-1}\left(\mathbf{x}_{\ell}^{b},\, K\right)_{\ell}\mathbf{M}\left(\mathbf{x}_{\ell-1}^{a},\, K\right)^{*}_{\ell,\,\ell-1}\cdot\mathbf{x}_{\ell-1}^{a} + \mathbf{H}_{\ell}^{T}\cdot\mathbf{R}_{\ell}^{-1}\cdot\mathbf{y}_{\ell}\right] \quad (17c)$$

Based on equation (17c), we denote by $\mathbf{x}_{\ell}^{a}(K)^{\infty}$ the analysis produced by using weights (3b) in the computation of (17a), and by $\mathbf{x}_{\ell}^{a}(K)^{2}$ the analysis obtained by employing weights (3c) in the computation of (17a).

Now, we are ready to test our proposed methods by using real-life meteorological data.

4 Experimental Results

To assess the accuracy of our method, we make use of the NCEP-DOE Reanalysis II dataset [5]. In this set, some physical variables such as air temperature T, relative humidity q, and wind components u and v are available four times daily, and those are the ones that we consider in our experiments. In this context, t_{ℓ} denotes the hour of the day in which snapshots are taken, $t_{\ell} \in \{0, 6, 12, 18\}$. Thus, we have four linear models:

- model $\mathbf{M}^{*}(\mathbf{x}_0, K)_{0,1}$ propagates a state $\mathbf{x}_0$ from hour 0 to hour 6,
- model $\mathbf{M}^{*}(\mathbf{x}_1, K)_{1,2}$ propagates a state $\mathbf{x}_1$ from hour 6 to hour 12,
- model $\mathbf{M}^{*}(\mathbf{x}_2, K)_{2,3}$ propagates a state $\mathbf{x}_2$ from hour 12 to hour 18, and
- model $\mathbf{M}^{*}(\mathbf{x}_3, K)_{3,0}$ propagates a state $\mathbf{x}_3$ from hour 18 to hour 0 (next day).

For each hour t_{ℓ}, by using the ensemble of snapshots $\mathbf{X}_{\ell}$, a $\widehat{\mathbf{B}}_{\ell}^{-1}$ matrix is estimated. A general structure of our piecewise linear model under these settings can be seen in figure 1.

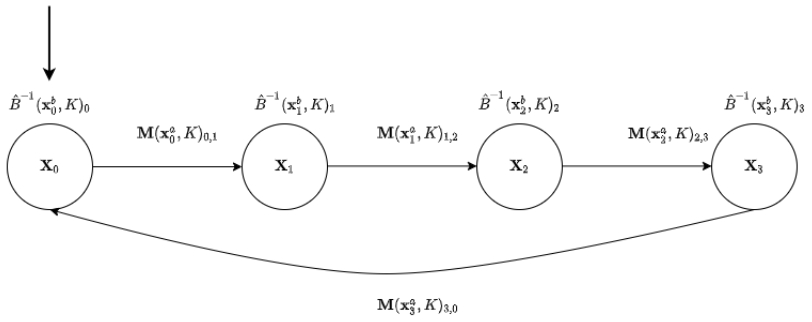


Fig. 1: Piecewise first order Markovian model for the NCEP-DOE Reanalysis II dataset.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77970-2_25

Additional settings are described below:

- for all physical variables, we consider snapshots on the surface level,
- we consider the number of nearest states $K \in \{10, 20, 50\}$,
- local linear models are computed by using the weights (3b) and (3c),
- the local models are trained with snapshots from January 1st to July 31rd, 2020,
- the forecast and data assimilation process are carried out between august 24, 2020 to august 27,2021,
- observations are taken every six hours (corresponding with the elapsed time between snapshots), these are simulated from the dataset by employing the following standard deviation for observation errors:
  - Temperature $1\,K^o$.
  - Zonal Wind Component $1\,m/s$.
  - Meridional Wind Component $1\,m/s$.
  - Specific Humidity $10^{-3}\,kg/kg$.

  ,
- the number of observations is only 5% of model components, these are randomly placed during assimilation steps,
- as measures of accuracies, we consider the L_2-norm of errors

$$\epsilon_k = \|\mathbf{x}_k^* - \mathbf{x}_k\|_2 \,, \tag{18}$$

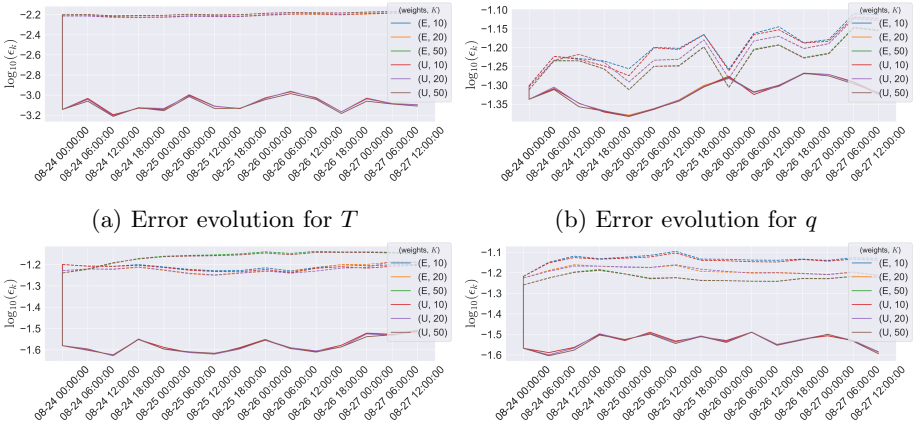
where $\mathbf{x}_k^*$ and $\mathbf{x}_k$ are the reference state (actual snapshot) and its approximation at time t_ℓ, respectively. $\mathbf{x}_k$ can be obtained via pure forecasts or analysis states. The Root-Mean-Square-Error (RMSE) measures, in average, the performance of a method within an assimilation window with M time spaced set of observations $\{\mathbf{y}_k\}_{k=1}^M$:

$$\overline{\epsilon} = \sqrt{\frac{1}{M} \cdot \sum_{k=1}^{M} \epsilon_k} \,. \tag{19}$$

4.1 Time Evolution of Errors

In figure 3, the L_2-norm of errors are shown for pure forecasts as well as analysis states. The results are shown for different values of K and the uniform weights (3b) (U) and the Euclidean ones (3c) (E). We let dashed and solid lines represent forecasts and analysis errors, respectively. As can be seen, the proposed models (3a) can produce accurate forecasts. For instance, error levels remain almost constant for the entire assimilation window for all model variables. This can be explained as follows: since our general model is piecewise linear, linear models can properly capture actual dynamics in time periods of six hours. Moreover, by injecting "current" information into the system, such forecasts can be improved. For instance, errors in model variables such as T, u, and v can improve over a magnitude, in some cases. Improvements can be seen in the q variable as well for the overall time window. Note that forecast errors are sensitive to the choices

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77970-2_25



(a) Error evolution for T

(b) Error evolution for q

(c) Error evolution for u

(d) Error evolution for v

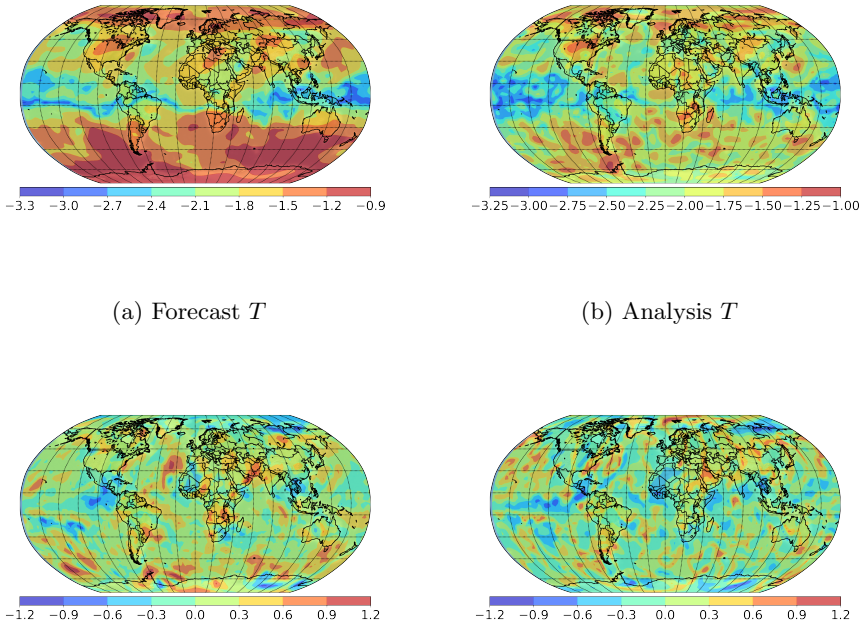
Fig. 2: Error evolution of model variables in the L_2 norm. Dashed lines denote pure forecasts while analysis solutions are shown in solid lines. Results are presented in the log-scale for easiness in reading.

of the number of neighbors K and the chosen distance (U or E). This is very common in the context of ML methods. Besides, those parameters are strictly related to the computation of (3a). For some variables, the differences in model trajectories for pure forecasts are evident. Nevertheless, the analysis solutions for all cases remain similar; their differences are imperceptible (recall that errors are in log-scale). This is very relevant since one of the current challenges in ML based methods is the tune of model parameters such as K. However, by using data assimilation (17c), the resulting states seem to be non-sensitive to the choices of K nor the weighing metric for distances. Of course, further research is needed to come up with a satisfactory answer to this.

4.2 Mean of Errors for the Assimilation Window

In figure 3, we show the average of L_2-norm of errors for some model variables. As can be seen, the assimilation of observations can improve the quality of forecasts. For instance, in the T variable, many dense regions owing to the accumulation of errors can be dissipated by using the information brought by observations of the system. Furthermore, the use of local information for computing the precision matrices (17b) allows for the dissipation of spurious correlation in errors of distant model components (in space). The use of local samples (snapshots) can mitigate the impact of multi-modal prior error distributions (which are commonly in highly non-linear models). In this manner, samples that potentially belong to another mode of the prior distribution are neglected during the estimation of prior error moments. Similar behavior can be noted for other model variables such as v.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77970-2_25



(a) Forecast T (b) Analysis T

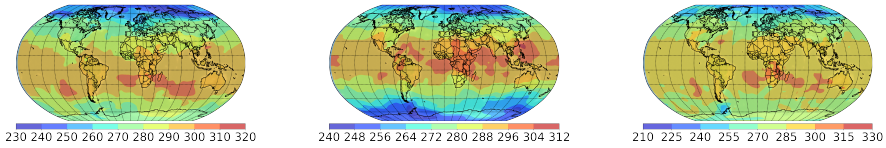
(c) Forecast v (d) Analysis v

Fig. 3: Mean of L_2-norm of errors in the log-scale for the assimilation window for $K = 10$ and the weighting metric U.

In figure 4, we show the initial background, analysis state, and reference snapshot for the variable T. This step is of interest since no actual information of the system has been injected into the numerical model. We can see some spurious waves near the south pole for the forecast state. For instance, a lower level of temperature than those of the actual state is reported. We can see that, by using only 5% of observations, the accuracy can be drastically improved. For instance, the temperature levels are adjusted similarly to those of the actual snapshot. This mainly obeys the implicit background error correlations captured in $\widehat{\mathbf{B}}_0^{-1}(\mathbf{x}_0{}^b, K)$: analysis increments are properly weighted as background error correlations are well estimated. This has a very important impact on the updating process of model components wherein no observations are available. In our settings, most of the Earth is unobserved (95%).

In the Table 1, we report the RMSE values for all model variables and all configurations. As we mentioned before, errors in forecasts are highly sensitive to ML parameters, as should be expected since these models rely on parameters

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77970-2_25



(a) Actual Snapshot (b) Forecast State (c) Analysis State

Fig. 4: Snapshots in the globe for the T variable by using Euclidean weights and $K = 10$.

such as K. Something that is very attractive is that we do not see high variations in errors, which is important since ML parameters can be hard to tune (i.e., parameter selection can vary from problem to problem). Hence, forecast errors are of similar magnitude regardless of the choice of K, for instance. On the other hand, we can see similar results of analysis errors for each configuration. Again, this obeys the fact that background error correlations are properly estimated during assimilation steps. Furthermore, the impact of multi-modal prior error distributions can be mitigated by considering similar snapshots to $\mathbf{x}_\ell^b$ during the estimation of $\widehat{\mathbf{B}}^{-1}(\mathbf{x}_\ell^b, K)_\ell$.

		Analysis			Forecast		
	K	10	20	50	10	20	50
Distance	Variables						
E (3c)	T	0.0288	0.0287	0.0283	0.0790	0.0786	0.0802
	q	0.2190	0.2192	0.2187	0.2528	0.2482	0.2427
	u	0.1629	0.1626	0.1625	0.2479	0.2454	0.2624
	v	0.1713	0.1712	0.1705	0.2709	0.2540	0.2442
U (3b)	T	0.0288	0.0287	0.0283	0.0790	0.0789	0.0803
	q	0.2190	0.2190	0.2187	0.2513	0.2481	0.2425
	u	0.1629	0.1626	0.1625	0.2466	0.2448	0.2616
	v	0.1712	0.1712	0.1705	0.2692	0.2540	0.2439

Table 1: Root-Mean-Square-Error values for the entire assimilation window, for all variables and different configuration of parameters.

5 Conclusions

In this paper, we propose a piecewise first-order Markovian model for the weather forecast. The proposed method employs linear models to mimic the behavior of

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77970-2_25

the system in time intervals. For each time interval, a linear model is built to forecast system states. The accuracy of the forecast can be improved by using Machine Learning models such as the K-Nearest-Neighbor regression. In this context, we employ weighing distances to compute linear models: the reciprocal of the uniform and the euclidean distances. Besides, sequential data assimilation can be exploited to inject real-time information of the system into our Markovian model. The estimation of background error correlation is performed by using a modified Cholesky decomposition and considering the K nearest snapshots to the forecast state. Experimental tests are performed by employing the National-Centers-for-Environmental-Prediction Department-of-Energy Reanalysis II dataset. The results reveal that numerical forecasts can be obtained within reasonable error magnitudes in the L_2 norm sense, they do not blow up, and even more, observations can improve forecasts by order of magnitudes, in some cases, for the entire assimilation window.

Acknowledgment

This work was supported by the Applied Math and Computer Science Laboratory (AML-CS) at Universidad del Norte, BAQ, COL.

References

1. Bickel, P.J., Levina, E., et al.: Covariance regularization by thresholding. The Annals of Statistics **36**(6), 2577–2604 (2008)
2. Bouttier, F., Courtier, P.: Data assimilation concepts and methods march 1999. Meteorological training course lecture series. ECMWF **718**, 59 (2002)
3. Evensen, G.: Data assimilation: the ensemble Kalman filter. Springer Science & Business Media (2009)
4. Houtekamer, P.L., Mitchell, H.L.: Ensemble kalman filtering. Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography **131**(613), 3269–3289 (2005)
5. Kanamitsu, M., Ebisuzaki, W., Woollen, J., Yang, S.K., Hnilo, J., Fiorino, M., Potter, G.: Ncep–doe amip-ii reanalysis (r-2). Bulletin of the American Meteorological Society **83**(11), 1631–1644 (2002)
6. Law, K., Stuart, A., Zygalakis, K.: Data assimilation. Cham, Switzerland: Springer **214** (2015)
7. Nino-Ruiz, E.D., Sandu, A., Deng, X.: An ensemble kalman filter implementation based on modified cholesky decomposition for inverse covariance matrix estimation. SIAM Journal on Scientific Computing **40**(2), A867–A886 (2018)
8. Nino-Ruiz, E.D., Sandu, A., Deng, X.: A parallel implementation of the ensemble kalman filter based on modified cholesky decomposition. Journal of Computational Science **36**, 100654 (2019)
9. Reichle, R.H.: Data assimilation methods in the earth sciences. Advances in water resources **31**(11), 1411–1418 (2008)
10. Richardson, L.F.: Weather prediction by numerical process. Cambridge university press (2007)

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77970-2_25

11. Tarantola, A.: Inverse problem theory and methods for model parameter estimation. SIAM (2005)
12. Zhang, S., Li, X., Zong, M., Zhu, X., Cheng, D.: Learning k for knn classification. ACM Transactions on Intelligent Systems and Technology (TIST) **8**(3), 1–19 (2017)

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77970-2_25