Integrating data assimilation with structurally equivariant spatial transformers: Physically consistent data-driven models for weather forecasting

Abstract

While recent years have seen an increase in interest to build data-driven models using deep learning techniques for seamless weather forecasting, the prediction horizon for skillful forecast remains inferior to numerical weather prediction (NWP) models. This can be attributed to the insufficient physics that is used as inputs to the deep learning models, inability of the models to perform high-quality spatio-temporal forecasts on physical fields, and the general challenges that exist due to the interacting scales of motion in turbulent flow. While in practice, operational weather models have data assimilation to correct the trajectory of the forecasts every 6 hours, the forward forecasting model, typically a high-resolution NWP makes the framework computationally expensive. In this paper, we show that a carefully chosen data assimilation scheme can be coupled to a deep learning based forecast model that allows the framework to maintain correct trajectories for several weeks. We ensure that the deep learning model is structurally equivariant (thus constraining key rotational features in the physics of the flow) and can perform skillful forecasts with only partial input (without the knowledge of the full physics or the equations of motions used in an NWP) while the data assimilation schemes incorporate noisy observations every 24 hours. The framework shows skillful predictions for multiple weeks starting from a noisy initial condition without losing trajectory of the forecasts on the Z500 field of the ECMWF Reanalysis 5 (ERA5) dataset. We conclude that such hybrid deep learning and data assimilation frameworks can enable better forecast performance for data-driven weather prediction and can be extended to operational-quality multi-ensemble probabilistic weather forecasts at a fraction of current computational cost.

1 Introduction

Data-driven models for prediction of different aspects of weather have received a lot of attention in recent times [1, 2, 3, 4]. These models, leveraging deep learning techniques, have advantages in terms of being computationally efficient, especially during inference mode (once training is done and weights are frozen) as compared to NWP models. Data-driven models, once trained on either high resolution climate model outputs or observed data, can also be used to perform parameterization of key small-scale physics that are often approximated in an ad-hoc fashion or ignored in practical low-resolution weather/clmate models [5, 6] . However, the current challenge in fully data-driven weather forecasting models is that it is outperformed by operational NWP models in terms of prediction horizon of forecasts. Yet, for practical purposes, these data-driven models can inexpensively be used to build very-large ensemble probabilistic forecasts that can come close to competing with NWPs at a fraction of the computational cost.

It must be noted that operational weather forecasts have data assimilation (DA) schemes integrated within them in order to ensure that forecast trajectories remain accurate. The DA scheme assimilates information from partial and noisy observations of the atmosphere in order to update the states

AI for Earth Sciences Workshop at NeurIPS 2020.

of the atmosphere predicted by the NWP model and thereby updates the initial condition to be used for the next forecast cycle. In this paper, we present a novel equivariance preserving spatial transformer based encoder-decoder model for weather prediction that integrates a DA scheme in order to produce accurate forecasts of the Z500 field (obtained from the ERA5 dataset) for several weeks. The equivariance preserving neural network can handle distortions in the physical input space in terms of rotation and stretching of large-scale waves in the atmosphere thus enabling it to extract features that are more physically relavant to the turbulent flow field that is expected to evolve data-drivenly. The advantage of using a deep learning model is the low computational cost associated with forward forecasting as compared to an NWP. Coupled with DA, data-driven models can remain competitive with operational forecasting since the DA scheme enables the data-driven model to recover back the trajectory it loses as forecasting lead time increases.

In this paper we address some of the key challenges in data-driven weather forecasting, namely, *physically constraining deep learning to capture large-scale rotational features of a turbulent flow field, seamless prediction with knowledge of only partial dynamics, and synergistic integration of data assimilation with deep learning based weather forecasting models.* Our contribution to this paper is three fold:

- We introduce an equivariance preserving deep spatial transformer module inside a convolutional encoder-decoder [7] network that enables the architecture to be physically consistent in terms of capturing large scale rotations of rossby-wave-like features in the global atmosphere and provide marginally better prediction skill.
- We perform seamless weather forecasting with the knowledge of partial dynamics, i.e., predicting the evolution of Z500 with only the knowledge of an initial noisy Z500 (training is also done on Z500 without the knowledge of any other key variables that are responsible for affecting the dynamics of the atmopshere, such as T2m, T850, and Z50).
- We integrate a sigma-point ensemble Kalman filter algorithm for data assimilation of a noisy observed Z500 every 24 hours of forecast with the data-driven model thereby ensuring that the data-driven forecasts remain skillful for weeks.

2 Framework and Physically Consistent Neural Architecture

2.1 Localization Network or Encoding Block

The architecture takes in an input snapshot of a noisy initial condition for Z500, $Z(t)^{32\times 64}$ (standard deviation of the gaussian white noise is $\sigma = 3 \times Z_\sigma$ where Z_σ is the standard deviation of time mean of Z from 1979-2015) and projects it onto a low dimensional encoding space via a standard convolutional encoding block. This encoding block performs three convolutions (without changing the spatial dimensions) and max-pooling which connects to a dense network with three layers. The encoded feature map connects into the spatial transformer module (see section 2.2). The convolutions inside the encoder block ensure periodic boundary conditions zonally by performing circular convolutions on each feature map inside the encoder block.

2.2 Spatial Transformer Module

The spatial transformer takes the feature map as the input and then applies an affine transformation $\Omega(\theta)$, where the parameters θ are learnt through backpropagation. One could use any other transformation but since the vortices rotate on the $x - y$ plane in our system, we have chosen an affine transformation so that that relative rotational and stretching features in the flow can be captured. The affine transformation results into transforming the original co-ordinate space x_i^o and y_i^o into x_i^s and y_i^s following

$$\begin{bmatrix} x_i^s \\ y_i^s \end{bmatrix} = \Omega(\theta) \begin{bmatrix} x_i^o \\ y_i^o \end{bmatrix} \tag{1}$$

The transformer module then applies a differentiable sampling kernel (a bi-linear interpolation kernel in this case) to the input feature map (denoted by F) at the locations given by x_i^s and y_i^s to produce the output feature map G. This can be written as

$$G_i = \sum_{n=0}^{n=32} \sum_{m=0}^{m=64} F_{nm} k(x_i^s - m; \Phi_x) k(y_i^s - n; \Phi_y) \tag{2}$$

Here $i \in [1, 2, \cdots 32 \times 64]$ and k is a generic sampling kernel with sampling parameters Φ_x and Φ_y. In our implementation k is a bi-linear interpolation kernel which is differentiable. This module ensures that the features extracted and thus the latent space encoded is equivariant and physically consistent by encoding the rotational features inherently present in the turbulent flow field.

2.3 Decoding Block

The decoding block is a standard series of deconvolution layers (convolution with zero-padded upsampling) that bring the encoded equivariant latent space back into its original dimension at time $t + \Delta t$ (Δt being 1 hour in this case) thus outputting $Z(t + \Delta t)^{32\times 64}$.

2.4 Data assimilation

Every 24 hours in this forecast cycle we implement a sigma point ensemble kalman filter with 2048 ensembles that assimilates a noisy observation of $Z(t)$ at every 24th hour. This enables the forecast to regain back skill owing to the new observation, albeit noisy, every 24th hour. Due to brevity, we do not discuss the implementation of the ensemble kalman filter (EnKF) which can be found in any standard textbook for data assimilation. Note that, the DA scheme is applied once the model is trained and deployed at inference stage. The new observations do not affect the trained weights of the network during inference. A schematic of the proposed framework is shown in Fig. 1.

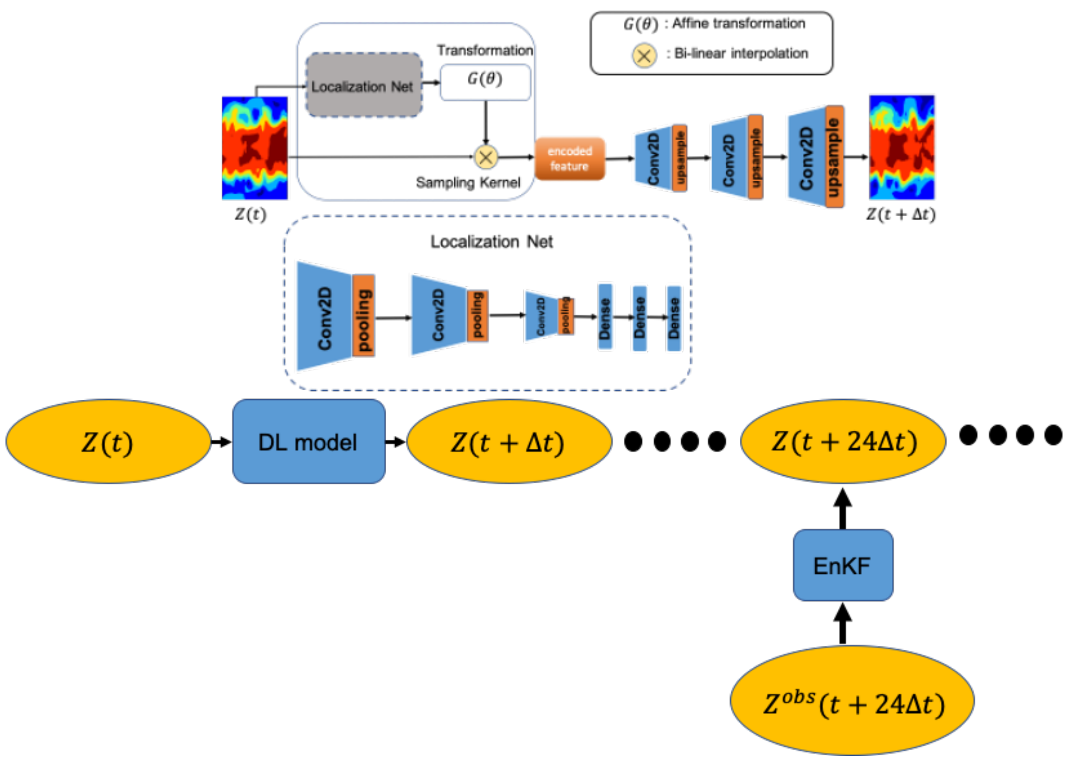

Figure 1: Equivariant spatial transformer architecture and the proposed framework to integrate the model with an EnKF DA scheme. During inference stage, the trained architecture performs a Δt integration of $Z(t)$ initialized with noise (see section 2.1) where $\Delta t = 1$ hour. Every 24 hours, a sigma point EnKF scheme assimilates information from a noisy observed $Z(t)$ such that the predicted trajectory remains closely correlated to the truth (more in section 2.6).

2.5 Training and Testing

We train the model on hourly Z500 data from the re-gridded ($5.625°$) ERA5 dataset between $1979 - 2014$ with input and labels being pairs of $(Z(t), Z(t + \Delta t))$ where $\Delta t = 1$ hour. The validation set is drawn from the year 2015 and the testing set is derived from the year of 2018.

2.6 Results and Discussions

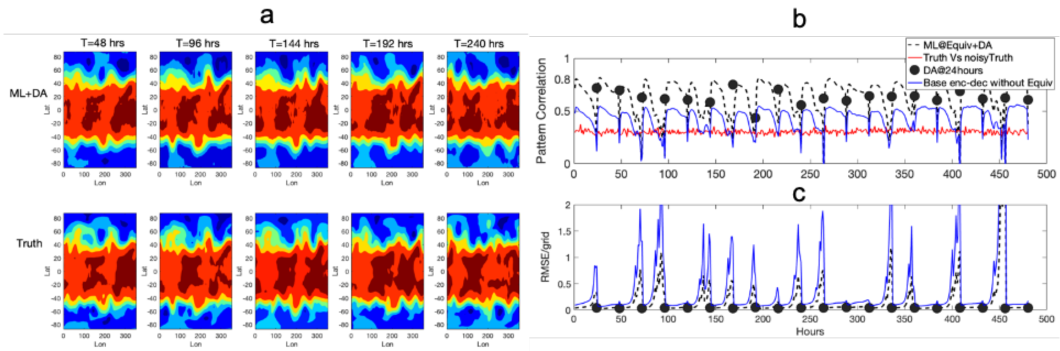

Figure 2: Forecast performance of the propsed framework. (a) Z500 patterns predicted by our framework upto 240 hours in the future starting from a noisy initial condition for Z500 in January 2018. (b) Pattern correlation of the obtained forecasts with the truth (without noise). Black dashed: proposed equivariant architecture with DA integration for upto 20 days, red solid: true observations with the noisy observations that are assimilated by the DA algorithm, and blue solid: baseline convolutional encoder-decoder with DA integration. (c) RMSE error per grid point for the same frameworks as (b). It is evident that the proposed equivariant spatial transformer proposed when integrated with DA outperforms a baseline encoder-decoder architecture. Note that, the initial condition has a large noise, since observations usually obtained from satellites has inherent noise associated with it.

We can see in Fig 2 (a) that the Z500 patterns forecasted remains accurate and physical multiple days into the future (shown upto 10 days). Large-scale rossby-wave-like structures are well captured along with most fine-scale structures on the jet stream. Figure 2(b) shows that the proposed architecture with physical constraints in the form of equivariance consistently shows better pattern correlations (close to 0.80 correlation coefficient) as compared to the baseline encoder-decoder architecture (close to 0.50 correlation coefficient) both of them integrated with DA during inference stage. In Fig. 2(c) we show RMSE per grid point of each of the Z500 patterns predicted and the conclusions remain the same as in the case of Fig. 2(b). For both the architectures (proposed and baseline), we have used the same noisy initial condition from January 2018. This promises us that equivariance as a key physical constraint generally improves the predictive capabilities of the data-driven model and DA ensures that the information incorporated from the noisy observations push the trajectory of the forecast to be consistently of high quality despite long lead time (~ 20 days).

We have chosen to use a large standard deviation for the noise in the observations to build a challenging test case. The correlation between the noisy truth (observations assimilated by DA) with the truth is low (correlation coefficient of 0.30). Yet, the forecasting performance of the proposed framework remains high with the DA step ensuring that the forecast keeps a physical trajectory close to the true one with an increase in lead days despite the absence of key dynamical fields in the input to the model.

We conclude that it is possible to perform data-driven prediction of certain features in weather with knowledge of partial dynamics, once key physical constraints can be enforced (e.g., rotational variations in large-scale vortices). DA can be integrated into such data-driven models thus alleviating the immense computational expense associated with NWP models. Synergistic integration of DA with equivariant deep learning forecasting architectures keeps a physical trajectory of forecasts even at long lead times. Further possibilities of such hybrid frameworks and different DA algorithms should be explored in the future for data-driven seamless weather forecasting.

References

[1] Sebastian Scher and Gabriele Messori. Weather and climate forecasting with neural networks: using general circulation models (GCMs) with different complexity as a study ground. *Geoscientific Model Development*, 12(7):2797–2809, 2019.

[2] Ashesh Chattopadhyay, Ebrahim Nabizadeh, and Pedram Hassanzadeh. Analog forecasting of extreme-causing weather patterns using deep learning. *Journal of Advances in Modeling Earth Systems*, 12(e2019MS001958), 2020.

[3] Ashesh Chattopadhyay, Pedram Hassanzadeh, and Devika Subramanian. Data-driven predictions of a multiscale lorenz 96 chaotic system using machine-learning methods: reservoir computing, artificial neural network, and long short-term memory network. *Nonlinear Processes in Geophysics*, 27(3):373–389, 2020.

[4] A Weyn, Jonathan, R Durran, Dale, and Rich Caruana. 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*, 10(8):2680–2693, 2019.

[5] Stephan Rasp. Online learning as a way to tackle instabilities and biases in neural network parameterizations. *arXiv preprint arXiv:1907.01351*, 2019.

[6] Ashesh Chattopadhyay, Adam Subel, and Pedram Hassanzadeh. Data-driven super-parameterization using deep learning: Experimentation with multi-scale lorenz 96 systems and transfer-learning. *arXiv preprint arXiv:2002.11167*, 2020.

[7] Ashesh Chattopadhyay, Mustafa Mustafa, Pedram Hassanzadeh, and Karthik Kashinath. Deep spatial transformers for autoregressive data-driven forecasting of geophysical turbulence. *EarthArxiv:10.31223/osf.io/cqmb2*, 2020.