Using Data Assimilation to Improve Data Driven Surrogate Models

Michael Goodliff[1,2] and Takemasa Miyoshi[1,2]
1: RIKEN Center for Computational Science, Japan
2. RIKEN Center for Interdisciplinary Theoretical and Mathematical Sciences (iTHEMS), Japan

Abstract:
Data-driven models (DDMs) are mathematical, statistical, or computational models built upon data, where patterns, relationships, or predictions are derived directly from the available information rather than through explicit instructions or rules defined by humans. These models are constructed by analysing large volumes of data to identify patterns, correlations, trends, and other statistical relationships. In areas such as numerical weather predictions (NWP), these DDMs are becoming increasingly popular with an aim to replace numerical models based on reanalysis data. Data assimilation (DA) is a process which combines observations from various sources with numerical models to improve the accuracy of predictions or simulations of a system's behaviour.

This presentation focuses on the application of DA methodologies in enhancing the precision and efficiency of DDM generation within computation models characterised by inherent observation error. The aim is to demonstrate the pivotal role that DA techniques can play in refining and optimising the process of DDM generation, thereby augmenting the accuracy and reliability of predictive models despite the presence of observational uncertainties.

Methodology:

Looping Algorithm: In figure 1, we outline our proposed algorithm to improve data-drive model generation using data assimilation. In this algorithm, we start with an imperfect model. This model is then used with data assimilation on perfect model observations. Using the analysis trajectory, we can generate an LSTM (LSTM gen 0) to be a better estimate of the system than the imperfect model. If we then repeat this step, by using the new LSTM instead of the numerical model, this produces a new LSTM (LSTM gen 1) which is, again, a more accurate representation of the perfect system. This algorithm can be looped *n* times (here we use *n=10*) to produce a machine learning model that is closer to the perfect system.

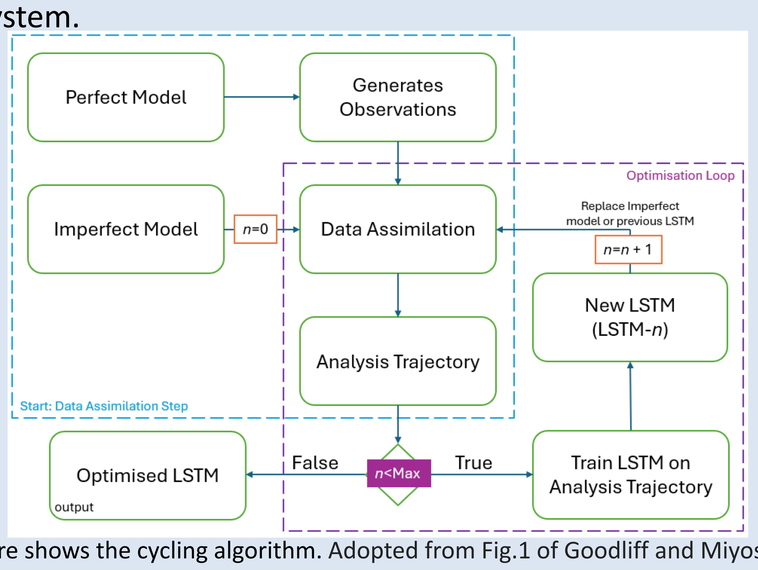


Figure 1. This figure shows the cycling algorithm. Adopted from Fig.1 of Goodliff and Miyoshi (2025).

Lorenz 63 Systems:

1) **Traditional Lorenz 63 Equations**
2) **"Coupled Chaotic" Lorenz 63 Model**

$$\frac{dx_0}{dt} = -\sigma(x_0 - x_1) + x_4$$
$$\frac{dx_1}{dt} = -\rho x_0 - x_1 - x_2 x_0 + x_3$$
$$\frac{dx_2}{dt} = x_0 x_1 - \beta x_2$$
$$\frac{dx_3}{dt} = -\omega x_4 - k(x_3 - x_3^*) - x_1$$
$$\frac{dx_4}{dt} = \omega(x_3 - x_3^*) - k x_4 - x_0$$

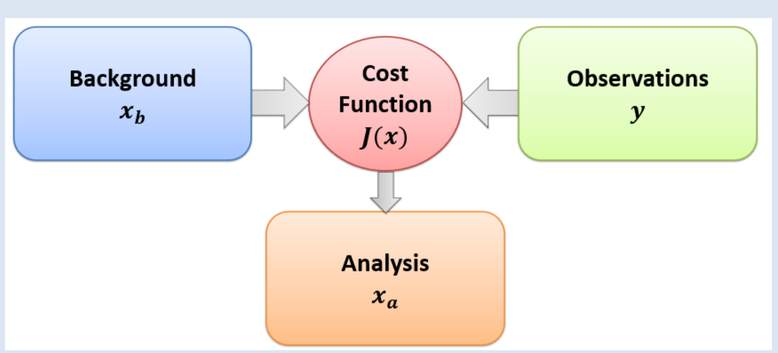
Square-Root Extended Kalman Filter (SR-EKF):

SR-EKF is a data assimilation method that recursively estimates the state of a nonlinear system by optimally combining model forecasts and observations while propagating the square root of the error covariance to preserve numerical stability. The prediction equations are:

$$\mathrm{x}_\mathrm{k}^\mathrm{f} = M_{\mathrm{k}-1}(\mathrm{x}_{\mathrm{k}-1}^\mathrm{a}) \qquad \mathrm{P}_\mathrm{k}^\mathrm{f} = \mathrm{F}_{\mathrm{k}-1}\mathrm{P}_{\mathrm{k}-1}^\mathrm{a}\mathrm{F}_{\mathrm{k}-1}^\mathrm{T} + \mathrm{Q},$$

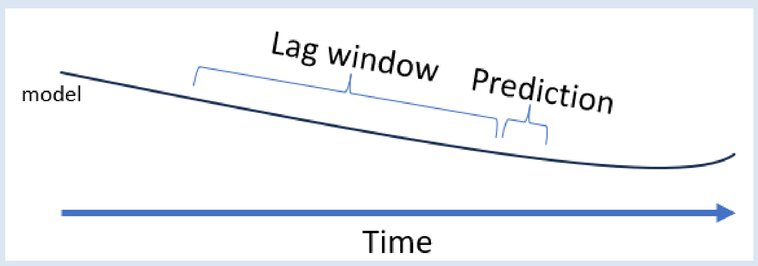
And the update equations are:

$$\mathrm{K}_\mathrm{k} = \mathrm{S}_\mathrm{k}^\mathrm{f}\left(\mathrm{H}_\mathrm{k}\mathrm{S}_\mathrm{k}^\mathrm{f}\right)^\mathrm{T}\left[\left(\mathrm{H}_\mathrm{k}\mathrm{S}_\mathrm{k}^\mathrm{f}\right)\left(\mathrm{H}_\mathrm{k}\mathrm{S}_\mathrm{k}^\mathrm{f}\right)^\mathrm{T} + \mathrm{R}\right]^{-1} \qquad \mathrm{x}_\mathrm{k}^\mathrm{a} = \mathrm{x}_\mathrm{k}^\mathrm{f} + \mathrm{K}_\mathrm{k}\left(\mathrm{y}_\mathrm{k} - \mathrm{H}_\mathrm{k}\left(\mathrm{x}_\mathrm{k}^\mathrm{f}\right)\right)$$



Long-Short Term Memory (LSTM):

LSTMs are a type of RNN capable of learning order dependence in sequence prediction problems. Using the past N data points in a time series, we can predict the next point.



Results:

Figure 2 illustrates the model error (RMSE) of our data-driven models (LSTMs) forecasts (1-4dt) compared to the imperfect model. As expected, the imperfect model exhibits the highest RMSE, indicating the lowest accuracy, while the LSTMs models become more accurate per iteration until convergence.

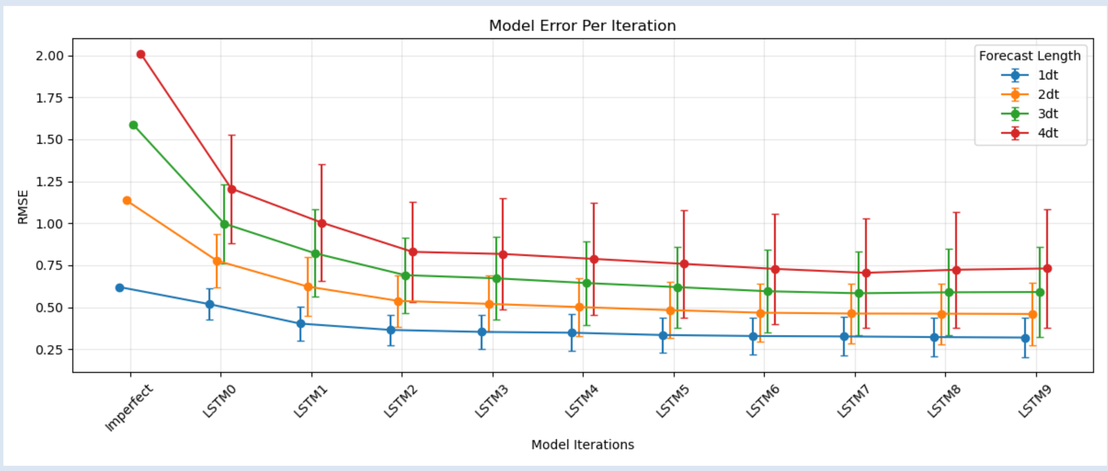


Figure 2. This figure shows the model error (RMSE) of our cycling algorithm on the Lorenz 63 model for a forecast of 1-4dt. As we cycle through the algorithm, each iteration improves the RMSE over the imperfect model. Adopted from Fig.2 of Goodliff and Miyoshi (2025).

Figure 3 shows the model error (RMSE) in space for the imperfect model, the first-generation LSTM-0, and the optimised LSTM. Here, we show that our optimised LSTM has a lower higher overall accuracy at all forecast lengths but struggles around the boundaries of the attractor at 4dt.

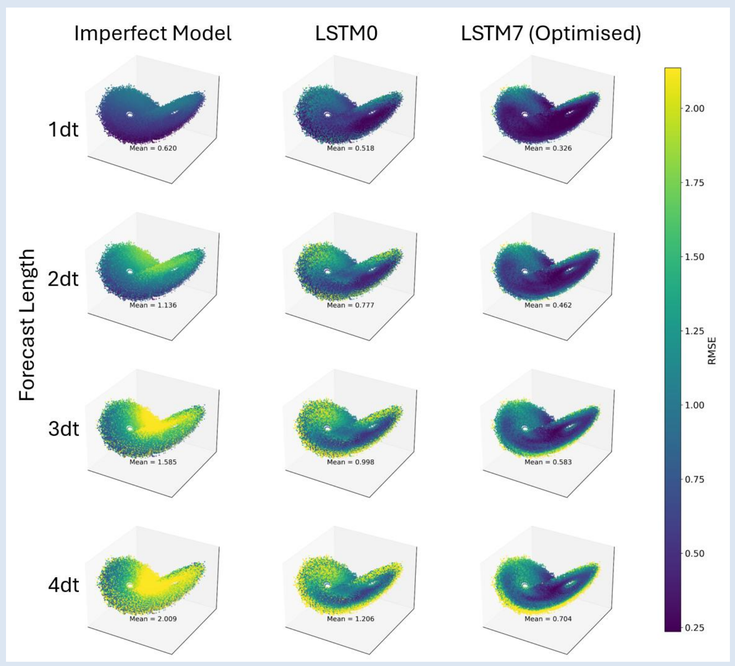


Figure 3. This plot illustrates the spatial distribution of model errors for 1–4*dt* forecasts. Adopted from Fig.4 of Goodliff and Miyoshi (2025).

Figure 4 shows how our imperfect model, initial LSTM (LSTM0) and optimal LSTM (LSTM7) change with varying levels of model error. This looks at $\beta = \{2.5, 2.8, 3.1, 3.4\}$, the imperfect model changes dramatically with higher model error, while the LSTMs stay relatively stable.

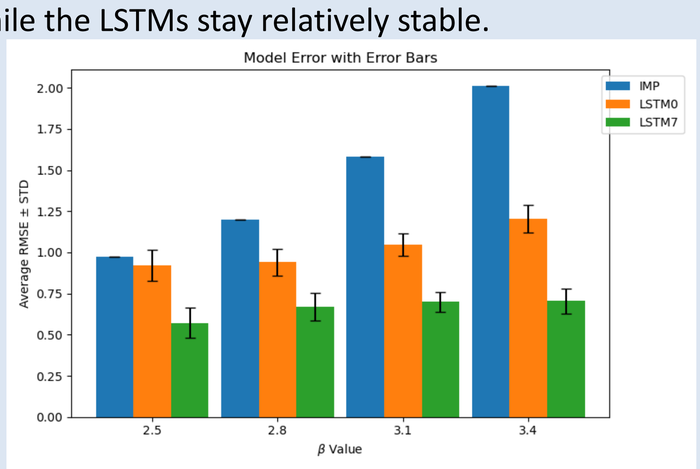


Figure 4. . This figure displays the model error (RMSE) for the imperfect model, initial LSTM (LSTM0) and the optimised LSTM (LSTM7).

References:
1. ZB Bouallègue, et al., The rise of data-driven weather forecasting: A first statistical assessment of machine learning–based weather forecasts in an operational-like context. Bull. Am. Meteorol. Soc. 105, E864 – E883 (2024).
2. J Brajard, A Carrassi, M Bocquet, L Bertino, Combining data assimilation and machine learning to emulate a dynamical model from sparse and noisy observations: a case study with the Lorenz 96 model. Geosci. Model. Dev. Discuss. 2019, 1–21 (2019).
3. E Lorenz, Deterministic Nonperiodic Flow. Journal of The Atmospheric Sciences 20, 130–141 (1963).
4. TN Palmer, Extended-range atmospheric prediction and the Lorenz model. Bull. Am. Meteorol. Soc. 74, 49 – 66 (1993).
5. Lorenc, A.C. Analysis methods for numerical weather prediction. *Quarterly Journal of the Royal Meteorological Society*, 112, 1177–1194 (1986).
6. M Goodliff and T Miyoshi, Using Data Assimilation to Improve Data-Driven Models. (Preprint) EGUsphere, doi.org/10.5194/egusphere-2025-933 (2025)