Dynamics of real-time forecasting failure and recovery due to data gaps

Sicheng Wu and Ruo-Qian Wang

Department of Civil and Environmental Engineering, Rutgers, The State University of New Jersey, Piscataway, NJ 08854

(Dated: September 9, 2022)

Real-time forecasting is important to the society. It uses continuous data streams to update forecasts for sustained accuracy. But the data source is vulnerable to attacks or accidents and the dynamics of forecasting failure and recovery due to data gaps is poorly understood. As the first systematic study, a Lorenz model-based forecasting system was disrupted with data gaps of various lengths and timing. The restart time of data assimilation is found to be the most important factor. The forecasting accuracy is found not returning to the original even long after the data assimilation recovery.

Introduction - Recent decades have witnessed a transition of forecasting systems (e.g. for weather and flood) from the traditional model-centric system to the cyber-physical system that integrates numerical models and monitoring sensor networks: the data from sensor networks is used for model calibration, validation, and data assimilation [1]. As this trend continues, monitoring networks are becoming increasingly important in key applications such as extreme weather preparation and warning, disaster responses, water supply and irrigation planning, energy, industry, recreational, and ecosystem water uses that are critical to national, state, tribal, and local economic well-being [2].

Such sensor-based monitoring-forecasting systems are being challenged by numerous environmental and social issues, causing data stream disruptions and threatening the system accuracy and reliability. External forcing, such as extreme weather, wildfire, vandalism, and telecommunication issues could compromise the availability and quality of data to harm the reliability of forecasts in important applications [3]. Unscheduled maintenance due to cyber-security attacks [4][5], lack of maintenance due to budget cuts [6] and shutdowns caused by large-scale health emergencies also continuously threaten the availability and reliability of the system.

The impact of the monitoring data stream disruptions on the accuracy of forecasting is poorly understood. The forecasting method incorporating real-time data is called data assimilation, which continuously integrates new measurements from monitoring networks to update the state of a deterministic forecasting model. The dominating method of data assimilation is Kalman Filter (KF), which was developed in the 1960s for optimal control of systems governed by linear equations, and its variants [7]. However, KF-based data assimilation schemes were designed without security in mind [8]. This loophole makes the monitoring-forecasting system vulnerable to accidents and targeted attacks. The wireless communication community developed a few strategies to enhance the resilience of KF-based systems [3][9].

To the authors’ knowledge, all the past studies focused on the accuracy of data transferring and controllability of robotic systems, and no systematic study has been performed to understand the fundamental physics behind the nonlinear system. This study is designed to fill the knowledge gap. Specifically, we aim at answering the following questions:

- How will the forecast error grow when the real-time data stream fails?
- Will the forecast error return to the level before the data gap after the real-time data stream recovers?
- Which factors determine the processes of the forecasting break down and recovery?

This Letter is targeted to use a classical theoretical model to study the dynamics of data stream failures and recovery.

Dynamics model and data assimilation - This study focuses on the classical chaos model developed by Lorenz [10], which represents the simplified atmospheric convection rolls that are sensitive to the initial conditions. The Lorenz model is a highly nonlinear model and has the typical chaotic behavior. It is worth noting that real forecasting systems are designed to be stable and using the Lorenz model to represent the real system is an exaggeration of the operational system behaviors. Nevertheless, we still select the Lorenz model in this study because of the convenience to trigger observable changes.

The governing equation of the Lorenz model can be written as:

\[
\begin{align*}
\frac{dx_1}{dt} &= \sigma(x_2 - x_1) \\
\frac{dx_2}{dt} &= x_1(\rho - x_3) - x_2 \\
\frac{dx_3}{dt} &= x_1x_2 - \beta x_3
\end{align*}
\]

(1)

in which \( \rho, \beta \) and \( \sigma \) are the model parameters consistent with the original configuration, i.e. \( \rho = 28, \beta = 8/3, \sigma = 10 \). This model is annotated as \( x(t + \Delta t) = f(x(t), \rho, \beta, \sigma) \).
An Ensemble Kalman Filter (EnKF) is used in this study to perform data assimilation. EnKF is a variant of the original KF, in which an ensemble of state members is used to estimate the covariance [1]. Specifically, the first step is to advance an ensemble of state vectors using the mechanics model by inserting the data of the ensemble \((\dot{x}_{i-1})\) into the Lorenz equations (Eq. 1) to calculate the next time step \(\dot{x}_i\). This time advancing continues until the assimilation time step \(t = t_a\) is reached, at which moment the state of the model is update following

\[
\dot{x}_i = (I_n - K_{ta} H_{ta}) \dot{x}_i + K_{ta} y_{ta}
\] (2)

where \(H_{ta}\) is the observation matrix, \(K_{ta}\) is the Kalman gain matrix estimated using the ensemble. Next, a re-sampling step is performed to regenerate an ensemble of new state vectors based on the estimated covariance to continue the next cycle of data assimilation. Data Assimilation Research Testbed (DART) [12] is used to implement the data assimilation. The details are listed in Table I.

| Numerical Method         | Runge-Kutta 2\textsuperscript{nd} order |
|--------------------------|----------------------------------------|
| Forecasting time step (\(\Delta t\)) | 0.001                                  |
| Assimilation time step (\(\Delta t_a\)) | 0.006                                  |
| Total assimilation steps | 10000                                   |
| Ensemble size            | 80                                      |

TABLE I. Setup of the data assimilations with DART.

A perfect run, \(X_0\), a reference run with data assimilation of continuous data, \(X_R\), and assimilated runs with data gaps, \(X_G\), are performed, where \(X = \{x(0), x(\Delta t), x(2\Delta t), \ldots\}\) is the time series of the modeling result. The perfect simulation \(X_0\) is used as the “ground truth” data, which is considered the accurate physics to measure the forecasting error. Because no model is perfect to capture all the details of the physics to make precise predictions, we built a perturbed Lorenz model by increasing \(\beta\) of the perfect model by a small value of \(10^{-14}\) to represent the systematic error of the forecasting model, i.e. \(f_P = f(x, \rho, \beta + 10^{-14}, \sigma)\). Synthetic observational data, \(X_{obs}\), is generated by adding a Gaussian random sampling process with a variance of 8.0 to \(x_{truth}\), i.e. \(X_{obs} = x_0 + \epsilon_0\), where \(\epsilon_0 \sim N(0, 8^2)\). The first data assimilation run, \(X_R\), was obtained by running \(f_P\) with the assimilation of \(X_{obs}\) for a total length of 60 time units (\(\Delta t = 0.006\)) without any data gap to serve as the reference. The second run, \(X_G\), was obtained with a similar procedure but a data gap was created by removing the observation data from \(T_{start}\) with different lengths, \(T_L\). Specifically, the procedure can be summarized in three phases:

1. “Normal” phase – the forecasting model, \(f_P\), was run to generate a time series of forecast by assimilating the synthetic observation \(X_{obs}\);

2. “Gap” phase – \(f_P\) was continued without assimilating the synthetic observation starting from \(T_{start}\);

3. “Recovery” phase – \(f_P\) was continued at \(T_{end} = T_{start} + T_L\) but with the assimilation of the synthetic observation \(X_{obs}\) again, which was designed to recover the observation data stream.

In this study, a series of experiments was conducted to generate \(X_G\) with \(T_{start}\) spreading from 3.3 to 9.3 (non-dimensional time) and \(T_{end}\) spreading from \(T_{start} + 0.6\) to 10.8.

The prediction errors are defined as the ensemble average of the square errors between the predictions, \(f_G(x)\), and the ground truth, \(f_0(x)\), for the reference (\(\epsilon_R\)) and the run with gap (\(\epsilon_G\)) respectively: [13]:

\[
\epsilon(t) = \frac{1}{N} \sum_{i=1}^{N} [x(t) - x_0(t)]^2, \quad (3)
\]

where \(x\) and \(x_0\) are respectively the prediction and ground truth data at \(t\), and the index \(i\) stands for each of the \(N\) ensembles (\(N=80\)).

The error departure is defined as the difference between \(\epsilon_G\) and \(\epsilon_R\):

\[
\Delta_G = \epsilon_G - \epsilon_R \quad \text{(4)}
\]

which measures how much the data gap impacts the data assimilation result.

Results - The typical behavior of the data assimilated Lorenz Model is demonstrated in Figure 1(a)–(d). The red and green dots represent the ensemble location of the “gap” and “reference” cases, \(X_G\) and \(X_R\). The ensemble was generated according to the covariance at the beginning of the simulation (Figure 1(a)). When the data gap begins, the ensemble was concentrated to the same place according to the small covariance (Figure 1(b)). In the Gap phase (Figure 1(c)), the ensemble of the reference (green dots) kept a narrow spread thanks to the data assimilation, but the ensemble of the “gap” case widely spread because no observation can be used to “correct” the model state. Once the data stream was restored, the spread of the ensemble shrank again and the “gap” case quickly approached the reference (Figure 1(d)). This process is quantified in Figure 1(e), where \(\epsilon_G\) and \(\epsilon_R\) are compared over the process. The errors of both cases were exactly the same until the beginning of the data gap. Then, \(\epsilon_G\) grew exponentially while \(\epsilon_R\) remained as in the “Normal” phase. In the recovery phase, \(\epsilon_G\) significantly dropped to approach \(\epsilon_R\) to match despite a small discrepancy.

A series of numerical experiments were performed with a range of \(T_{start}\) and \(T_L\). We found that the slope of the exponential growth, which is known as the Lyapunov exponent (denoted \(\lambda_0\)), is always around 0.87 for \(\epsilon_G\).
in the “Gap” phase using linear fitting. The maximum value of \( \epsilon_G \) through all the cases was found at the level of \( 10^3 \).

\[ T_R = T_{rec} - T_{end} \]  \hspace{1cm} (5)

\( T_R \) is found increasing with \( T_L \) when \( T_L \) is small and being constant with \( T_{end} \) when \( T_L \) is large (Figure 2c). Regarding the position of the data assimilation restarts, \( T_R \) is found increasing with \( \epsilon_G(T_{end}) \) when \( \epsilon_G(T_{end}) \) is small, but it kept constant when \( \epsilon_G(T_{end}) \) is greater than \( 10^{1.5} \) (Figure 2d). These observation could inform a strategy to recover the data stream for improved forecasting accuracy and more discussion is detailed at the end of this Letter.

An interesting observation is made in Figure 1(e) that despite \( \epsilon_G \) quickly returned to the reference once the data assimilation recovered, it never exactly restored to \( \epsilon_R \) even long after \( T_{end} \). This phenomenon is better reflected in the error departure \( \Delta_G \) for a sample case with a short \( T_L \) shown in Figure 3. It shows in the Recovery phase \( \epsilon_G \) in most time exceeded \( \epsilon_R \) (\( \Delta_G > 0 \)) while it did become lower than \( \epsilon_R \) in several instances (i.e. \( \Delta_G < 0 \)). Three strong sporadic departure peaked around of 25% of the maximum \( \Delta_G \) during the data gap are observed. The first peak is observed during the recovery of \( \epsilon_G \) shortly after \( T_{end} \). The two subsequent sporadic departure peaks occurred after \( \Delta_G \) returned to near-zero at \( T = 12.0 \). Some departure peaks of smaller magnitude also occurred after the two strong departure peaks. This demonstrates that the data gap had a permanent disruption to the state of the Lorenz model and a “hot” recovery of the data assimilation could not fully restore the accuracy of the

\( \epsilon_G \) quickly decreased towards \( \epsilon_R \) after the recovery of the data assimilation. To quantify the process, we use the minimum linear slope of \( \epsilon_G \) based on the moving linear fitting starting 5 data points after \( T_{end} \) with a step of 3 data points until the forecasting recovery ends at \( T_{rec} \) – the first time the forecasting error returns to the reference error, i.e. when \( \Delta_G \) reaches 75% of the \( \Delta_G \) at \( T_{end} + 20 \). The most negative slope or the minimum Lyapunov exponent is denoted \( \lambda_1 \) to characterize the recovery rate of the system.

\( \lambda_1 \) is plotted against \( T_L \) in Figure 2(a) to find important factors that impact it. This figure shows that when \( T_L \) was relatively small, \( \lambda_1 \) decreases with \( T_L \), while when \( T_L \) is large enough \( \lambda_1 \) weakly depends on \( T_L \) but \( T_{end} \) plays a more significant role. A further analysis shows that \( \lambda_1 \) declines exponentially with the forecasting error at the beginning of the data stream recovery, \( \epsilon_G(T_{end}) \) (Figure 2b). This indicates that the greater the forecasting error, more quickly the forecasting accuracy restores.

\[ T_{end} = 8.4 \quad T_{end} = 9.6 \quad T_{end} = 10.8 \]

\( T_{end} \) is found increasing with \( T_L \) when \( T_L \) is small and being constant with \( T_{end} \) when \( T_L \) is large (Figure 2c). Regarding the position of the data assimilation restarts, \( T_R \) is found increasing with \( \epsilon_G(T_{end}) \) when \( \epsilon_G(T_{end}) \) is small, but it kept constant when \( \epsilon_G(T_{end}) \) is greater than \( 10^{1.5} \) (Figure 2d). These observation could inform a strategy to recover the data stream for improved forecasting accuracy and more discussion is detailed at the end of this Letter.

An interesting observation is made in Figure 1(e) that despite \( \epsilon_G \) quickly returned to the reference once the data assimilation recovered, it never exactly restored to \( \epsilon_R \) even long after \( T_{end} \). This phenomenon is better reflected in the error departure \( \Delta_G \) for a sample case with a short \( T_L \) shown in Figure 3. It shows in the Recovery phase \( \epsilon_G \) in most time exceeded \( \epsilon_R \) (\( \Delta_G > 0 \)) while it did become lower than \( \epsilon_R \) in several instances (i.e. \( \Delta_G < 0 \)). Three strong sporadic departure peaked around of 25% of the maximum \( \Delta_G \) during the data gap are observed. The first peak is observed during the recovery of \( \epsilon_G \) shortly after \( T_{end} \). The two subsequent sporadic departure peaks occurred after \( \Delta_G \) returned to near-zero at \( T = 12.0 \). Some departure peaks of smaller magnitude also occurred after the two strong departure peaks. This demonstrates that the data gap had a permanent disruption to the state of the Lorenz model and a “hot” recovery of the data assimilation could not fully restore the accuracy of the
original forecast. In average, the mean error of \( \Delta G \) after \( T_R \) is 4.94\% \pm 3.50\% based on the standard deviation, with an maximum of 12.7\% across all cases.

FIG. 3. A sample of time series of (a) \( \epsilon_G \) and (b) \( \Delta G \) with \( T_{\text{start}} = 5.7 \) and \( T_{\text{end}} = 7.8 \) to show the forecasting errors long after the restart of the data assimilation.

Discussion - The Lyapunov exponent of around 0.9 in the Gap phase is consistent with previous studies, e.g., [14]. This is expected because without data assimilation the system became a perturbed Lorenz system and the typical bifurcation occurred. This further indicates that a nonlinear, chaotic system will exponentially accumulate the forecasting error if the real-time data stream is lost and the rate of the error accumulation is dependent on the stability of the mechanistic model.

The strange attractors in the phase diagram limited \( \epsilon_G \) to the range about 10\(^{3.2} \). It took about 800 time steps for any case to reach this boundary and the error stopped growing after this saturation (Figure 3(d)). This observation is consistent with the known characteristics of the system, i.e. the predictability of the Lorenz model is completely lost at around 1000 time steps [19].

The saturation point is also shown in the analysis of \( \lambda_1 \) and \( T_R \). At the right end of Figure 2(b) and (d), the error reached the saturation point so that the recovery processes are similar. While the recovery process before the saturation depends on how much error accumulates when the data assimilation restores. As an analogy, the system takes a “spring” style to recover – the greater the accumulated error, the steeper and longer the recovery process takes. This can be demonstrated in Figure 4 with the same \( T_{\text{end}} \), \( \lambda_1 \) decreases and \( T_R \) increases with the accumulated error until a threshold that the recovery processes converged.

These discoveries could inform forecasting operators to restore a real forecasting system after a data stream failure. First, in a bounded forecasting system, the recovery of the data assimilation will follow the same pattern when the data gap is long enough, i.e. the saturation point is reached. Second, if the data gap is not long enough to reach the error saturation, the operators should fix the system as soon as possible to reduce the full recovery time. Third, because the forecasting error is still present long after the data streams are restored, the restored forecasting system may be different from the perfect forecasting that has continuous data streams. To fully restore the forecasting accuracy, it is necessary to rerun the data assimilation with continuous and consistent data streams.

In summary, this Letter is probably the first endeavor to systematically examine the dynamics of failure and recovery of real-time forecasting systems due to data gaps. Data gaps of various lengths and timing were created in a Lorenz model-based data assimilation system. We discovered that the forecast error grows exponentially with a Lyapunov exponent of around 0.9 because after losing the data assimilation the system becomes a typical chaotic model and experiences a bifurcation. After the observational data recovers, the prediction error declines with a rate faster than the growth of error and the recovery rate and time are positively correlated with the length of the data gap until saturation. Forecast error is still observed even long after the recovery of the data assimilation.
[1] K. Beven, Environmental modelling: an uncertain future? (CRC press, 2018).

[2] R. R. Mason Jr and B. A. Weiger, Stream Gaging and Flood Forecasting: A partnership of the US Geological Survey and the National Weather Service, Tech. Rep. (US Dept. of the Interior, US Geological Survey: National Oceanic and . . . , 1995).

[3] USGS, USGS Working to Restore Streamgages — U.S. Geological Survey (2018).

[4] M. Gruss, Noaa admits to cyberattack on satellite data networks (2014).

[5] G. L. Dodaro, Urgent Actions Are Needed to Address Cybersecurity Challenges Facing the Nation, Tech. Rep. GAO-18-645T (United States Government Accountability Office, 2018).

[6] L. Lundquist, Usgs stream gauges slowly return to service after nationwide failure (2018).

[7] J. A. Vrugt, C. G. Diks, H. V. Gupta, W. Bouten, and J. M. Verstraten, Improved treatment of uncertainty in hydrologic modeling: Combining the strengths of global optimization and data assimilation, Water resources research 41 (2005).

[8] D. Wagner, Resilient aggregation in sensor networks, in Proceedings of the 2nd ACM workshop on Security of ad hoc and sensor networks (2004) pp. 78–87.

[9] S. Nashimoto, D. Suzuki, T. Sugawara, and K. Sakiyama, Sensor con-fusion: Defeating kalman filter in signal injection attack, in Proceedings of the 2018 on Asia Conference on Computer and Communications Security ASI-ACCS ’18 (Association for Computing Machinery, New York, NY, USA, 2018) p. 511–524.

[10] E. N. Lorenz, Deterministic nonperiodic flow, Journal of atmospheric sciences 20, 130 (1963).

[11] M. Katzfuss, J. R. Stroud, and C. K. Wikle, Understanding the ensemble kalman filter, The American Statistician 70, 350 (2016).

[12] J. Anderson, T. Hoar, K. Raeder, H. Liu, N. Collins, R. Torn, and A. Avellano, The data assimilation research testbed: A community facility, Bulletin of the American Meteorological Society 90, 1283 (2009)

[13] A. Atencia and I. Zawadzki, Analogs on the lorenz attractor and ensemble spread, Monthly Weather Review 145, 1381 (2017).

[14] J. M. Nese, J. A. Dutton, and R. Wells, Calculated attractor dimensions for low-order spectral models, Journal of Atmospheric Sciences 44, 1950 (1987).