Partition of Analysis and Forecast Error Variance into Growing and Decaying Components

Jie Feng[1,2], Zoltan Toth[1], Malaquias Peña[3], and Jing Zhang[1]

[1] *Global Systems Division, ESRL/OAR/NOAA, Boulder, CO, USA*

[2] *School of Meteorology, University of Oklahoma, Norman, OK, USA*

[3]*Department of Civil and Environmental Engineering, University of Connecticut, Storrs, CT, USA*

Submitted to *Q. J. Royal Meteorol. Soc.*
June 14, 2019
Resubmitted to *Q. J. Royal Meteorol. Soc.*
Nov 5, 2019
Resubmitted to *Q. J. Royal Meteorol. Soc.*
Dec 19, 2019

* *Corresponding author address:*

Dr. Jie Feng

Address: School of Meteorology, University of Oklahoma,

120 David Boren Blvd. Norman, OK, 73072

E-mail: jie.feng@ou.edu

ABSTRACT

Due to the scarcity of and errors in observations, direct measurements of errors in Numerical Weather Prediction (NWP) analyses and forecasts with respect to nature (i.e., "true" error) are lacking. Peña and Toth (2014) introduced an inverse method called SAFE-I where true errors are (i) theoretically assumed to follow exponential error growth, and (ii) estimated from the perceived errors (i.e., forecast minus verifying analysis) that they affect. While decaying or neutral errors, by definition will not have a significant impact on longer range forecast errors, they can still accumulate in, and negatively influence NWP data assimilation–forecast cycles.

In a new, generalized version of the inverse method (SAFE-II), analysis and forecast error variance is decomposed into exponentially growing and decaying components, assuming they are independent as they comprise of vectors from the leading and trailing ends of the Lyapunov spectrum, respectively. SAFE-II uses the initial variance and decay rate associated with non-growing perturbations to describe and estimate their behavior.

The assumptions behind SAFE-II are first validated in a simulated environment. SAFE-II is then applied to estimate the error variance in both simulated and operational analyses/forecast environment. Perceived error measurements are found to be statistically consistent (at the 95% significance level) with the SAFE-II error behavior model, which offers a more accurate description of error variance than SAFE-I that neglects decaying errors. At various levels and for different variables, decaying errors

are found to constitute up to 60 % of the total analysis error variance, much of which decay during the first 12-18 hours of forecast integrations.

Keywords: uncertainty of analysis, forecast verification, error estimation, data assimilation, ensemble forecasts

1. Introduction

Due to the intermittency of, and errors in available observations, the true state of the atmosphere, however alluring it is, remains unknown. The state of the atmosphere is estimated using data assimilation (DA, current state, or analysis) and Numerical Weather Prediction tools (NWP, future states or forecasts). Both the assessment and improvement[1] of the quality of DA and NWP tools and products depend on reliable estimates of analysis and forecast error variance. In most studies, such errors are estimated with the variance between NWP analysis or forecast states that are being evaluated and verifying observations or NWP analysis fields (in case of forecast verification). Since errors in some verifying observations or analysis fields are of comparable magnitude to those in analysis or short-range forecast fields that are being evaluated, such an approach is convoluted and yields questionable results.

Peña and Toth (2014, PT14) introduced a method hereafter called Statistical Analysis and Forecast Error (SAFE-I) estimation that relates the measured *perceived forecast error variance* (forecast minus verifying analysis) to *true error variance* (forecast minus reality interpreted on the model grid). SAFE-I is independent of any assumptions used in analysis or forecast systems. The measured perceived forecast error variance is modeled by several unknowns. To reduce the number of unknowns in the statistical estimation process, it uses prior knowledge about the evolution of errors in

[1] For example, the reliable specification of analysis error variances offers a reference for the rescaling of initial ensemble perturbations (Toth and Kalnay, 1997; Molteni et al., 1996; Wei et al., 2008). Also, the accurate quantification of short-range forecast error variances can orient the tuning of background forecast error covariance in DA (Fisher, 1996; Whitaker et al., 2008).

48 analysis–forecast systems. The unknown parameters are estimated via the minimization
49 of the difference between the sample mean (e.g., over a season) of measured and
50 modeled (via the unknown parameters) perceived error variance. Feng et al. (2017)
51 extended the application of SAFE-I from area mean to pointwise error estimation and
52 quantified the spatial distribution of analysis and short-range forecast error variance at
53 a 95% confidence level.

54 For simplicity, SAFE-I assumes that in short range (i.e., out to 2 or 3 days) synoptic
55 scale forecasts all analysis errors grow at a close to exponential rate. Analysis errors are
56 therefore assessed in a "growing equivalent" sense. The effect of non-growing analysis
57 errors, if any, will implicitly manifest in modified estimates of the growing error
58 component. In the presence of a significant level of decaying analysis errors, this may
59 lead to an overestimation of initial growing error variance, and an underestimation of
60 the growth rate.

61 Analysis fields are a weighted sum of observations and NWP first guess forecast
62 fields. It is generally accepted that NWP analyses contain both random or decaying, and
63 dynamically conditioned, growing errors (Toth and Kalnay, 1993; 1997; Houtekamer et
64 al., 2005; Buizza et al., 2005; Wei et al., 2008; Peña et al., 2010). The former generally
65 signifies a lack of dynamical balance in analysis fields. These errors are believed to
66 originate from errors in observations (e.g., Hunt et al., 2007; Stewart et al., 2013), or
67 statistical DA approximations[2], hence can be considered random from a model

[2] Examples include the use of "covariance localization" in ensemble Kalman filters (EnKF) for the reduction of spurious long-distance covariances (Houtekamer and Mitchell, 2001). Such schemes may introduce an imbalance among different variables (Mitchell et al., 2002; Lorenc, 2003). The use of

68 dynamics point of view. Therefore, these errors project onto the stable (or decaying)
69 manifold of the system (Toth and Kalnay, 1997; Kalnay, 2003). Growing errors
70 originate from amplifying errors in first guess forecasts, projecting onto the unstable
71 (or growing) subspace (Pires et al., 1996; Toth and Kalnay, 1997; Kalnay, 2003;
72 Trevisan and Uboldi, 2004; Feng et al., 2018). As Pires et al (1996) showed, improved
73 DA techniques lead to a reduction of the proportion of errors that decay in the overall
74 analysis error.

75 When decaying errors are present in the analysis, over a transient period the overall
76 error may either decay or exhibit slower than exponential growth due to the rapid
77 collapse of random errors (e.g., Vannitsem and Nicolis, 1994; Trevisan and Legnani,
78 1995; Houtekamer et al., 2005; Palatella et al., 2013). Such a transient period is
79 followed by exponential error growth[3], characteristic of the system's dynamics
80 associated with the leading local Lyapunov vectors (Toth and Kalnay, 1997; Kalnay,
81 2003; Snyder and Hamill, 2003; Ding and Li, 2007; Li and Ding, 2011; Feng et al.,
82 2014).

83 Forecast errors also display a transitional decaying phase in Observing System
84 Simulation Experiments (OSSEs) where true error is directly measurable (see, e.g.,
85 Privé and Errico, 2013a). When initial perturbations are dynamically less conditioned
86 (i.e., have significant projection on the stable manifold due to, e.g. the addition of

incomplete balance constraints may also leave gravity waves in the analysis that appear as noise to hydrostatic models (Huang and Lynch, 1993; Kleist et al., 2009).

[3] In nonlinear systems, as the level of error becomes comparable to the size of the attractor, nonlinear interactions moderate exponential growth (Lorenz, 1982; Dalcher and Kalnay, 1987).

87 simulated observational noise), the ensemble spread may also exhibit transitional
88 behavior (e.g., Houtekamer et al., 2005; Hamill and Whitaker, 2011).

89 Decaying components of analysis error or perturbation variance rapidly disappear
90 during the initial phase of forecast integrations (typically in less than a day). But their
91 accurate estimation can (i) improve the accuracy of the analysis and short-range forecast
92 error variance estimation; (ii) diagnose the effectiveness of DA schemes (in the spirit
93 of Pires et al., 1996); and (iii) provide guidance as to the appropriate level of growing,
94 dynamically conditioned perturbations (as opposed to quickly disappearing noise) in
95 initial ensemble perturbation generation methods. In particular, the diagnosis of
96 decaying errors is a prerequisite for their reduction and for making analysis fields
97 dynamically more balanced.

98 This study is based on the recognition that analysis errors generally project onto
99 the full spectrum of local Lyapunov vectors (LLVs; Wolf et al., 1985; Legras and
100 Vautard, 1996), from the fastest growing to the fastest decaying directions (Toth and
101 Kalnay, 1997; Vannitsem and Nicolis, 1997; Hamill et al., 2002; Kalnay, 2003; Ding et
102 al., 2017; Feng et al., 2018). This is because each analysis step introduces some noise
103 into the analysis field, randomly projecting onto the full spectrum of directions in the
104 phase space. The forecast step amplifies dynamically growing error patterns while
105 dissipating errors in other directions, thus rotating the overall error toward the growing
106 subspace. Such potentially complex error behavior is approximated here by assuming
107 that the total error variance is the sum of two orthogonal error components (SAFE-II).
108 The first component is exponentially growing, characterized by the leading Lyapunov

vector (Lorenz, 1996; Toth and Kalnay, 1997; Ziehmann et al., 2000; Kalnay, 2003; Feng et al., 2014), estimated by SAFE-I, while the other component introduced here is exponentially decaying, considered as a composite of errors across all the neutral and trailing LVs.

The modeling of the decaying errors in SAFE-II is introduced in Section 2. Section 3 describes the Global Forecast System (GFS) that is used operationally at the National Centers for Environmental Prediction (NCEP), in which SAFE-II will be tested. The SAFE-II assumptions are validated in a GFS-based Observing System Simulation Experiment (OSSE) environment (Cucurull et al., 2017) where "ground truth" is known exactly (Section 4). Experimental SAFE-II results from both simulated and operational systems, including a comparison with SAFE-I output, are presented and analyzed in Section 5, followed by preliminary conclusions in Section 6 and discussion in Section 7.

2. Methodology

a. Statistical Analysis and Forecast Error estimation algorithm (SAFE-I)

Let **F**, **A**, and **T** denote the forecast, analysis, and true state of reality, all valid at the same time and interpolated onto a common model grid. The true ($\mathbf{x}_i$) and perceived ($\mathbf{f}_i$) errors in an $i \cdot \Delta t$ lead time forecast (where Δt is the length of the DA cycle) are then defined as:

$$\mathbf{x}_i = \mathbf{F}_i - \mathbf{T}_i \ , \tag{1}$$

$$\mathbf{f}_i = \mathbf{F}_i - \mathbf{A}_i \ . \tag{2}$$

Since the true state of reality is not known exactly, the true error is not measurable. For each lead time, PT14 introduces the following relationship between the true analysis and forecast error variances and the perceived forecast error variance measurements:

$$f_i^2 = x_0^2 + x_i^2 - 2\rho_i \cdot x_0 \cdot x_i, \quad (3)$$

where f_i^2, x_0^2 and x_i^2 are the spatial and temporal mean of error variance corresponding with $\mathbf{f}_i$, $\mathbf{x}_0$, and $\mathbf{x}_i$, and ρ_i is the sample mean correlation between $\mathbf{x}_0$ and $\mathbf{x}_i$. The unknown parameters are estimated by minimizing the difference between the measured (f_i^2) and the modeled ($\widehat{f_i^2}$) perceived error variance in Eqs. like (3).

Note that the number of unknowns in a series of Eqs. like (3) exceeds the number of measured quantities. Here we follow PT14 and use a simplifying setup as well as prior knowledge about error growth and DA (in the form of several assumptions, see Table 1) to dramatically reduce the number of unknowns in a series of Eqs. like (3).

Simplifying setup. As in SAFE-I, forecasts are verified against analysis fields from the same DA-forecast system that is used for the initialization of the forecasts:

$$\mathbf{F}_0 = \mathbf{A}_0 \quad (4)$$

We opt to use analysis fields instead of observations as a proxy for reality as, by design, they have a lower error. Choosing verifying analysis fields from the same system that initializes the forecasts reduces the number of unknowns, potentially reducing errors in their statistical estimation.

Assumption 1: Model error. In this study, we focus on extratropical forecast variables verified against analysis fields that represent natural processes at the model's spatiotemporal resolution. For simplicity, under these conditions we assume that model

error is negligible. In case total forecast error can be explained purely through the amplification of initial errors, the assumption will be considered validated. For other (e.g., tropical) variables or for processes not well resolved by the model (e.g., parameter or truncation errors), the model error can be explicitly represented as an additional term in Eq. 5 below (see, e.g., PT14, and Vannitsem and Toth, 2002, or Nicolis et al., 2009, respectively).

Assumption 2: Error evolution. Error variance in short-range forecasts of complex systems evolve exponentially and therefore can be described simply by two unknown parameters - the initial analysis error size x_0, and the exponential growth rate α (Lorenz, 1963):

$$x_i^2 = x_0^2 \cdot e^{i \cdot \Delta t \cdot \alpha} . \quad (5)$$

If necessary, (5) can be augmented to represent the effect of nonlinear saturation (i.e., replace the exponential relationship with the logistic function), or model related errors (PT14).

Assumption 3: Data impact on analysis. PT14 recognized that the repeated insertion of new observational information in successive DA-forecast cycles results in the progressive decorrelation of true error in a freely evolving forecast from the true error in verifying analyses valid at the same time. Assuming that a statistically similar amount of observational information is ingested in each DA cycle, the error decorrelation follows a power law relationship:

$$\rho_i = \rho_1^i . \quad (6)$$

174 ρ_1 and ρ_i in Eq. 6 indicate the angular extent to which the error in the latest analysis is
175 rotated from the error in the first guess (Δt = 6 hours for a typical DA cycle in global
176 forecast systems) or from earlier initialized longer range forecasts all valid at the time
177 of the analysis, respectively, due to one (or multiple) introduction(s) of observational
178 information. The simplicity of the data impact relationship in Eq. 6 is because the error
179 that is assumed to be comprised of leading LLVs, whether present in a DA-forecast
180 cycle or in a "free" longer range forecast, develop similarly in a quasi-linear fashion,
181 over the same (or very similar) time evolving flow.

182 With the relationships in Eqs. (4)-(6), the number of unknowns is significantly
183 reduced, and the short-range perceived error variance can be simulated with only three
184 unknowns (x_0, α, ρ_1):

185 $$\widehat{f_i^2} = x_0^2 + x_0^2 \cdot e^{i \cdot \Delta t \cdot \alpha} - 2\rho_1^i \cdot x_0^2 \cdot \sqrt{e^{i \cdot \Delta t \cdot \alpha}}. \quad (7)$$

186 The unknown parameters are then estimated by minimizing the cost function:

187 $$J = \max(|f_i^2 - \widehat{f_i^2}| \cdot w_i^{-1}), \quad (8)$$

188 where w_i^{-1} is the weight on the fitted perceived error variance at lead time $i \cdot \Delta t$, and $|\bullet|$
189 represents the absolute value. The minimization is carried out using the limited-memory
190 Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) algorithm (Byrd et al., 1995). The
191 choice of the L_∞ norm ($\max(\bullet)$; i.e. infinite norm) is motivated by a desire to get a good
192 fit over the entire range of lead times (PT14). Simulated perceived error values on the
193 right side of Eq. (7) are expected to match their measurement counterparts only within
194 the sampling uncertainty of the latter which is given by the Standard Error of the Mean
195 (SEM). For further details on SEM and w_i^{-1}, see Appendix A.

b. Decomposition of analysis and forecast errors

As mentioned earlier, analysis errors are generally assumed to project with varying power on the full spectrum of LLVs, from the fastest growing to fastest decaying vectors (Toth and Kalnay, 1997; Vannitsem and Nicolis, 1997; Hamill et al., 2002; Kalnay, 2003; Feng et al., 2018). Recognizing that true error variance at longer lead times is dominated by the fastest growing components of the total error (since decaying errors diminish early on), SAFE-I uses a most economical, 1-dimensional model to describe error evolution constrained in the subspace of the leading (i.e., fastest growing) LLVs.

Such a model, however, cannot describe the transitional behavior arising early on in a forecast due to neutral or decaying analysis errors. To assess the behavior of decaying errors and to enhance the accuracy of growing error variance estimates, here we propose a generalization of the SAFE-I algorithm. While SAFE-I assumes all errors are confined in the subspace of the leading LLVs and grow exponentially, the new method called SAFE-II introduces a second, exponentially *decaying* component orthogonal to the growing direction, accounting for all non-growing errors.

The total analysis error variance in SAFE-II is thus described as the sum of the growing and decaying components:

$$x_0^2 = g_0^2 + d_0^2, \tag{9}$$

where g_0^2 and d_0^2 are the initial growing and decaying error variances, respectively. In the forecast phase, the growing component expands exponentially, while the decaying component shrinks exponentially, yielding the following sum for the total true forecast error variance (*Assumption 2* behind SAFE-II):

$$x_i^2 = g_0^2 \cdot e^{i \cdot \Delta t \cdot \alpha} + d_0^2 \cdot e^{i \cdot \Delta t \cdot \beta}, \quad (10)$$

where β is a negative value representing the exponential decay rate. The transitional behavior of the total error (solid) due to the vanishing decaying errors (dotted) is illustrated in Fig. 1. Following the initial transitional period during which most of the decaying errors disappear, the total error follows the evolution of the exponential component (dashed line in Fig. 1). By substituting x_0^2 and x_i^2 in Eq. (7) with Eqs. (9) and (10), the perceived error variance simulated with the two additional SAFE-II parameters (d_0 and β) can be written as:

$$\widehat{f_i^2} = g_0^2 + d_0^2 + g_0^2 \cdot e^{i \cdot \Delta t \cdot \alpha} + d_0^2 \cdot e^{i \cdot \Delta t \cdot \beta} - 2\rho_1^i \cdot \sqrt{g_0^2 + d_0^2} \cdot \sqrt{g_0^2 \cdot e^{i \cdot \Delta t \cdot \alpha} + d_0^2 \cdot e^{i \cdot \Delta t \cdot \beta}}, \quad (11)$$

c. Use of additional measurements

Given the challenge of estimating two extra parameters compared to SAFE-I, we explored whether additional measurements beyond perceived errors could be used for the reduction of uncertainty in SAFE-II parameter estimation. Lagged Forecast Differences (i.e. the differences between two different lead time forecasts valid at the same time, hereafter LFD) is one such measurable quantity. In Fig. 2, $\mathbf{F}_{i\text{-}1}$ and $\mathbf{F}_i$ are two such forecasts (with lead times of $(i-1)\cdot\Delta t$ and $i\cdot\Delta t$), while **T** and **A** denote the true and analyzed states, respectively, all valid at the same time. In triangle $\mathbf{TF}_{i\text{-}1}\mathbf{F}_i$ (blue dotted lines), the LFD variance between $\mathbf{F}_{i\text{-}1}$ and $\mathbf{F}_i$ ($f_{i\text{-}1,\,i}^2$, red solid line) can be expressed as:

$$f_{i\text{-}1,\,i}^2 = x_{i\text{-}1}^2 + x_i^2 - 2\rho_{i\text{-}1,i} \cdot x_{i\text{-}1} \cdot x_i, \quad (12)$$

where $\rho_{i\text{-}1,i}$ is the correlation between $\mathbf{TF}_{i\text{-}1}$ and $\mathbf{TF}_i$ (blue dotted lines).

To enable the use of LFD measurements in SAFE-II without the introduction of $\rho_{i\text{-}1,i}$ as an additional unknown parameter, we introduced three additional assumptions.

Assumption 4: Relationship between true and perceived error variance. We first note that the correlation $\gamma_{i\text{-}1,i}$ between $\mathbf{AF}_{i\text{-}1}$ and $\mathbf{AF}_i$ (blue solid lines in Fig. 2) can be readily calculated from perceived error measurements. We further note that while true forecast error variance grows exponentially as a function of lead time, analysis error variance x_0^2 remains the same. Therefore, at sufficiently long lead times, the true and perceived errors become similar in magnitude:

$$f_{i\text{-}1}^2 \sim x_{i\text{-}1}^2 \text{ , and} \tag{13a}$$

$$f_i^2 \sim x_i^2, \tag{13b}$$

and therefore correlation $\rho_{i\text{-}1,i}$ can be well approximated by the measurable quantity $\gamma_{i\text{-}1,i}$. In this study we assume that the perceived and true forecast error variances become sufficiently similar at 2.25 (i=9) and 2.5 days (i=10) lead time, assuring that $\rho_{9,10} \approx \gamma_{9,10}$.

Assumption 5: Transient period. As further simplifications, we also assume that any transient error behavior subsides within 24 hours[4].

Assumption 6: Divergence rate of model trajectories. We also assume that the model, by reasonably capturing natural instabilities, reproduces the chaotic divergence of trajectories of the model and nature. Therefore, beyond 24-hour lead time, the three sides of triangle $\mathbf{TF}_{i\text{-}1}\mathbf{F}_i$ grow at the same pace, corresponding to the dynamically sustainable growth rate of the errors between model and reality (i.e. parameter α). It

[4] While true forecast error is largely unaffected by decaying errors beyond 24 hours lead time, the perceived error remains affected by decaying errors present in the verifying analyses, allowing for the estimation of decaying parameters in SAFE-II.

follows that correlations $\rho_{i-1,i}$ remain approximately unchanged beyond 24 hours ($i \geq 5$) and equal to $\rho_{9,10}$. *Assumptions 1-6* are summarized in Table 1 and their validity will be investigated in section 4 in an OSSE environment.

Based on the above assumptions and substituting x_i^2 in Eq. (12) with Eq. (10), the evolution of LFD between 1 to 2.5 days lead times can be modeled with only two parameters, g_0 and α, that are also used in the simulation of perceived forecast error variance (cf. Eq. (11)):

$$\widehat{f^2_{i-1,i}} = g_0^2 \cdot e^{(i-1)\cdot\Delta t\cdot\alpha} + g_0^2 \cdot e^{i\cdot\Delta t\cdot\alpha} - 2\gamma_{9,10} \cdot g_0^2 \sqrt{e^{(i-1)\cdot\Delta t\cdot\alpha} \cdot e^{i\cdot\Delta t\cdot\alpha}}, \ (i=5, 6, \ldots, 10). \quad (14)$$

To distinguish between the lead times of perceived error and LFD measurements and associated weights, the index i is replaced with j in Eq. (14) before an LFD term is incorporated into the SAFE-II cost function of Eq. (8):

$$J = \max\left(|f_i^2 - \widehat{f_i^2}| \cdot w_i^{-1}\right) + \max\left(\left|f^2_{j-1,j} - \widehat{f^2_{j-1,j}}\right| \cdot w^{-1}_{j-1,j}\right) \quad (i = 1, 2, \ldots 10\ ;\ j = 5, 6, \ldots 10) \quad (15)$$

Since the simulation of LFD (Eq. (14)) does not involve decaying errors, LFD variance measurements, if desired, can also be incorporated into the cost function of SAFE-I.

3. Data sets used

SAFE-II will be applied to estimate true error variance in GFS analyses and forecasts, first in a simulated (OSSE), then in a realistic operational forecast environment at NCEP. 6-, 12-, …, 60-hour perceived error measurements will be calculated over the extratropical Northern Hemisphere (NH; 30° - 90°N) on a $1^\circ \times 1^\circ$ regular latitude/longitude grid. A cosine weight of latitude is used when calculating the area mean error variance so as to avoid undue weights on data from higher latitudes.

The choices of the spatial domain and lead time range minimize the effects of nonlinearities (Gilmour et al, 2001) or model related errors (Orrell et al., 2001).

a. OSSE data

SAFE-II assumptions will be validated (Section 4) and estimates evaluated (Section 5) using the OSSE setup[5] described in detail by Appendix B and Cucurull et al., (2017). The three variables used are zonal wind (U), temperature (T), and geopotential height (GH). For the OSSE data used in this study, 6-hourly analyses are used corresponding with lower boundary conditions between 3 July and 26 August 2005, with 7-day forecasts initialized only at every 00Z. The SAFE-II cost function (Eq. (15)) is therefore modified to use 24-, instead of 6-hour lagged forecasts:

$$J = \max\left(\left|f_i^2 - \widehat{f_i^2}\right| \cdot w_i^{-1}\right) + \max\left(\left|f_{j\text{-}4,j}^2 - \widehat{f_{j\text{-}4,j}^2}\right| \cdot w_{j\text{-}4,j}^{-1}\right), (i = 1, 2, \ldots 10\,;\, j = 8, 9, 10), (16)$$

b. Operational GFS data

In Section 5, SAFE-II will also be used to assess analysis and forecast error variance in the operational, T1534, 64-level resolution GFS system (Yang, 2016). Analysis/forecast data are sampled every 6 hours and cover the 1 Sep - 30 Nov 2015 period.

Note that as mentioned before, the OSSE experiments introduced in Section 3.a use an earlier version of the NCEP NWP system (e.g., without a hybrid DA). Therefore, SAFE-II error estimates from the OSSE and operational environments cannot be directly compared.

[5] Short of having access to data from OSSE, analysis and forecast errors could also be simulated with perturbed fields from an ensemble of analyses and forecasts (Houtekamer et al., 2005; Feng et al., 2017).

4. Validation of SAFE-II assumptions

An OSSE environment offers an ideal ground for the evaluation of the assumptions behind SAFE-II since not only the perceived error, but various characteristics of the true forecast error (e.g., error variance and the correlation between analysis and forecast errors) are also directly measurable.

a. Basic assumptions

A key assumption (*Assumption 2* in Section 2.a and 2.b) states that the true forecast error variance can be considered as a sum of exponentially growing and decaying error components (Eq. (10)). In an OSSE environment, we can directly assess the validity of *Assumption 2* by fitting the error evolution relationship in Eq. (10) to the time mean of *true error variance measurements*, through minimizing the following cost function:

$$J = \max(\left| x_i^2 - \widehat{x_i^2} \right| \cdot \omega_i^{-1}),\ (i = 0, 1, \ldots 10), \tag{17}$$

where x_i^2 and $\widehat{x_i^2}$ are the measured and modeled true forecast error variances, respectively, and ω_i^{-1} is the weight related to the SEM-based sampling error (refer to Appendix A) of the measured true error variances.

Fig. 3 shows the sample (time) mean of directly measured true analysis and forecast error variance, along with 95% confidence intervals reflecting the effect of sampling errors, and a simulated error variance curve fitted to the sample mean of the measurements using the error decomposition of Eq. (10), as a function of 0-60 hours lead time, for three selected variables. All simulated values fall within the 95% confidence intervals, indicating that *Assumption 2* about the decomposition of forecast errors (Section 2.b) is consistent with the experimental data.

323 The four estimated parameters with SAFE-II are listed in Table 2 (the first 4 values
324 in "Fit SAFE-II" rows). These will be used as a reference for ground truth in the
325 evaluation of SAFE-II estimates in Section 5. The performance of SAFE-I is also shown
326 in Table 2. As expected, when decaying errors are absent (see variable T in Table 2), the
327 two versions of SAFE identify the same exponential error growth (with identical fitted
328 x_0^2 and α values). In the presence of decaying analysis errors (variables U and GH), the
329 SAFE-I error growth model still offers a statistically acceptable fit except for the
330 estimated total analysis error variance of GH. The good fitting of exponential growing
331 forecast error also justifies that *Assumption 1* can be considered to be valid. However,
332 with its additional two parameters, SAFE-II provides a considerably improved
333 simulation of the analysis error compared to SAFE-I (2-3%, instead of 8-13% deviation
334 from the reference measured true error, see corresponding numbers in parentheses in
335 Table 2).

336 When decaying errors are present (variables U and GH) but not considered, SAFE-
337 I arrives at higher initial growing error variance and lower growth rate estimates than
338 SAFE-II (Table 2). SAFE-II finds the largest proportion of decaying errors in GH (about
339 24.8% of the total analysis error variance), followed by U (11.2%). As an example, Fig.
340 4 illustrates the evolution of the estimated growing, decaying and total error variance
341 for GH by SAFE-II from 0 to 1.5 days. It is analogous to the behavior of growing and
342 decaying components of true forecast error variance in the schematic figure (Fig. 1).
343 Variables U has qualitatively similar error evolution. It is consistent for the variables
344 that decay is so fast that within 24 hours the percentage of decaying errors drops below

1% of the total forecast error variance. This validates *Assumption 5* in Section 2.c. The growth rate of LFD variance (the rightmost column in Table 2) has only up to 3% deviation from that of the true forecast error variance for all variables (Table 2) which indicates *Assumption 6* is reliable.

Another key assumption (*Assumption 3* in Section 2.a) states that the correlation between true analysis and forecast errors (ρ_i) exponentially decays with increasing lead time (Eq. (6)). In an OSSE environment, "ground truth" correlation values ρ_i can be diagnosed from true error variance measurements using a transformed version of Eq. (3):

$$\rho_i = (x_0^2 + x_i^2 - f_i^2) / (2 \cdot x_0 \cdot x_i). \quad (18)$$

To test *Assumption 3*, we simulate ρ_i with the exponential decorrelation relationship of Eq. (6) and then fit the simulated curve to the time mean of the diagnosed quantities (i.e., ground truth from Eq. (18)) by minimizing a cost function analogous to Eq. (17).

The results in Fig. 5 reveal a reasonable correspondence between sample-based mean and simulated values of the correlation between analysis and forecast errors. For the two model prognostic variables (U and T), the fitted values of ρ_i are within the 95% sampling error interval of its sample-based mean values throughout the first 2 days. This indicates that the exponential degradation of ρ_1 (*Assumption 3*) is consistent with the experimental data. Returning to Fig. 3 we observe that due to the relatively high correlation between analysis and short-range forecast errors (see Fig. 5), perceived errors for the model variables are significantly lower than the true errors (Fig. 3). At 6-hour lead time, for example, the perceived error measurements for U and T provide a

2-3-fold underestimate of the true error variance. This will be further discussed in the context of the operational forecast system in Section 5.a.2.

Note that the third variable shown in Fig. 5, GH, is not a directly analyzed variable; rather, it is derived from analysis control variables. Simulated ρ values for GH are nevertheless consistent with the ground truth, albeit only at and beyond 12 hours lead time. The deviation of 6-hour ρ is possibly due to some random noise or bias introduced in the calculation of GH (e.g., a particular discretization of the hydrostatic equation; Wee et al., 2012) in the OSSE that makes the ρ assumption invalid.

b. LFD-Related Assumptions

Recall from Section 2.c that correlation $\rho_{9,10}$ between lagged true forecast errors $\mathbf{x}_9$ and $\mathbf{x}_{10}$ in cost function (Eq. (14)) is specified by the correlation $\gamma_{9,10}$ between lagged perceived errors $\mathbf{f}_9$ and $\mathbf{f}_{10}$, valid at the same time. Since in an OSSE environment both angles are directly measurable, the accuracy of approximating correlation "ρ" with "γ" can be tested. Fig. 6 displays ρ and γ as a function of lead time with forecasts lagged 24 hours apart (since the OSSE forecasts are available only once, instead of four times per day) for variables U, T, and GH at 500 hPa height.

At 36/60 hour lead time, the correlation between lagged true and perceived errors differs less than 0.025. This can be explained by the small differences between true and perceived error variances at and beyond 36-hour lead time (see Fig. 3), validating *Assumption 4* (Section 2.c), and thus $\rho_{9,10} \approx \gamma_{9,10}$.

The correlation between lagged true errors in Fig. 6 exhibits less than 0.01 variations beyond 24/48 hours lead time. This indicates that once transient decay

subsides, triangles $\mathbf{TF}_{i-1}\mathbf{F}_i$ are approximately similar. This effectively validates *Assumption 6* about the close similarity of exponential expansion or growth rates in the attractor of a model and that of a model trajectory diverged from reality.

5 Assessment of error variance in OSSE and operational analysis/forecast systems

In this section, the SAFE-II algorithm described in Section 2 will be used to estimate GSI - GFS true analysis and forecast error variances. The estimates will be based on measurements of perceived error and LFD variances, first from an OSSE environment, then from the operational NCEP system.

a. Error variance in selected variables

1) OSSE Environment

In Section 4.a, error decomposition Eq. (10) was fitted directly to the time mean of true error variance measurements from an OSSE experiment. Here we will proceed as if we did not know reality and use perceived error variance measurements from the same OSSE analysis/forecast system to estimate the true error variance. In the error estimation experiments reported here, the truth will be used only in the evaluation of the results.

The quality of these practical estimates will be assessed by comparing them with a characterization of true growing (g_0^2) and decaying (d_0^2) error variances and their amplification and decay rates (α and β). As mentioned in Section 4.a, the fitted parameter values of SAFE-II in Table 2 are used as reference values for the estimates presented here. Along with these reference values, Table 3 shows the SAFE-I and

411 SAFE-II error parameter estimates for 500 hPa analysis variables U and T for the OSSE
412 experiments. Estimates of GH are strongly influenced by the deviation of ρ at 6 hours
413 (see discussion on Fig. 5(c)) and thus are not shown. SAFE-II estimates of growing
414 error variance and growing rate are closer to the reference values than the SAFE-I
415 results, except for T where the results are identical since no decaying component is
416 identified by SAFE-II. SAFE-II estimates are within 5-10% of the reference values for
417 growing error variance and within 2% for their amplification rate.

418 Since decaying errors shrink fast and practically disappear within 24 hours (see
419 Section 4.a), their estimation is especially challenging: only the first few perceived error
420 variance measurement points provide meaningful information about their behavior.
421 LFD measurements used in the cost function (see Eq. (14)) are no help with the
422 estimation of decaying parameters as they constrain only the estimation of the growing
423 parameters. SAFE-II decaying error variance and decay rate estimates for model
424 variable U contain relatively large, nearly 25% and 50% deviation from their reference
425 values, respectively. The total error variance estimate for U is more accurate wth SAFE-
426 II than SAFE-I (1% versus 9% error), though both are statistically reliable at the 95%
427 confidence level. Their estimates of error correlation are similar.

428 *2) NCEP operational analysis/forecast system*

429 The main results of the study are visualized in Fig. 7. 200 hPa (a) U, (b) T, and (c)
430 GH variables in the NCEP operational system are chosen for demonstration as decaying
431 errors constitute a sizable portion of analysis error at this level (see Section 5.b). Fig. 7
432 shows the perceived error variance measurements (black open circles) with a 95%

433 confidence level sampling uncertainty (black vertical bars) as a function of lead time.
434 The corresponding simulated perceived error variance (thin black curve), and the
435 estimated total (thick black curve), growing (red) and decaying (blue) error variance
436 are also shown.

437 For all variables and at all lead times, the simulated perceived error variance falls
438 within the 95% SEM uncertainty intervals of the perceived error measurements,
439 indicating that the SAFE-II error behavior model is consistent with the experimental
440 measurement data. The results in Fig. 7 confirm the error behavior indicated by the
441 schematic Fig. 1. After a relatively short, 12-18 hour transitional period during which
442 the decaying error component vanishes, the total error assumes an exponential growth.

443 Fig. 7 confirms a finding from the OSSE experiments (Section 4.a) that the
444 conventional measure of forecast performance, perceived error variance, can be a rather
445 poor estimate of the true short-range forecast error variance. At 6-hour lead time, for
446 example, the perceived error measurements provide a 3-4-fold underestimate of the true
447 error variance similarly as in the OSSE environment (see Fig. 3). The discrepancy is
448 due to the fact that perceived error measurements do not reflect the presence of error in
449 the verifying analysis fields that is relatively highly correlated with the error in the
450 forecasts. It is not until 2 days lead time that the deviation of simulated or measured
451 perceived error variance from the true error variance drops below 5% of the true error
452 variance. The use of perceived error as an estimate of true forecast error thus leads to
453 an underestimation of error variance and an overestimation of error growth (PT14). The
454 overestimation of error growth when using perceived error variances are used may

455 partially explain the apparent lack of sufficient spread and perceived deficiency in
456 perturbation growth in most ensemble systems studied (see, e.g., Buizza et al., 2005),
457 as well as the difference between "external" (i.e., verified against analyses of the
458 atmosphere) vs. "internal" (verified against another model forecast) predictability and
459 error growth noted by Lorenz (1982) and a series of follow-on studies.

460 *3) Comparison with results from OSSE analysis / forecast system*

461 Table 4 summarizes the NCEP operational forecast system results for 200 hPa
462 height variables displayed in Fig. 7. For an easy comparison with results from the error
463 evolution of an OSSE experiment in Section 4.a, results for 500 hPa height variables
464 are also shown in Table 4. Comparing SAFE-II estimates from Table 3 with 500 hPa
465 estimates from Table 4, we first note that both the total and growing analysis errors
466 appear to be severely underestimated by the OSSE system for U and GH. This may be
467 the result of tuning OSSE error variances to match operational perceived error variances
468 that, as noted above, are significantly lower than true error variances.

469 Interestingly, error growth rate estimates for the NCEP operational system verified
470 against operational analyses from September - November 2015 (Table 4), and against
471 an ECMWF high resolution simulation with July-August lower boundary forcing
472 (OSSE nature run, Table 3) display less than 3% difference for the two model variables
473 U and T at the 500 hPa height level. This may be an indication that when properly
474 assessed, external (model vs. reality) and internal (model vs. model) error growth, after
475 all, may be rather similar. These results are also consistent with *Assumption 6* in Section
476 2.c.

Analysis vs forecast error correlations are found slightly (about 0.03) higher in the operational system for all variables (cf. ρ_1 in Tables 3 and 4). For U, the OSSE analysis also contains a larger decaying component. The use of an improved hybrid DA scheme and increased model resolution in the operational vs. the OSSE setup, as well as the addition of too much noise in the simulation of observational error in the OSSE system may both contribute to the correlation and decaying error results above.

Table 4 also lists error variance and other estimated parameters for GH. While the GH analysis vs forecast error correlation in Table 4 appears to be similar to or only slightly higher than those for the other variables, both the growth rate and the percentage of decaying error is markedly higher for GH than for the other variables. The latter result is qualitatively consistent with OSSE results in Table 2. Note that GH is not a GFS model or GSI analysis variable but rather is derived from model prognostic variables including temperature, surface pressure, and humidity (Houtekamer et al. 2005) through the hydrostatic equation (Grell et al. 1995). When the hydrostatic equation is integrated from the model surface to the top of the model to calculate GH, independent random error present in the prognostic variables may lead to a higher level of noise (i.e., decaying error) in GH compared with model prognostic variables. As for the GH growth rate, it corresponds to an error doubling time of 1.26 days, below Simmons et al (1995)'s 1.5 days estimate for 500 hPa height. It is not clear why the GH growth rate is significantly higher than that for the model variables.

b. Vertical profile of analysis error variance

In this section, SAFE-II is used to estimate the vertical distribution of error variance from 1000 to 100 hPa for the GFS-GSI operational forecast system. SAFE-I estimates are also given for a comparison.

1) Fitting error

As described in Section 2.a, a critical part of SAFE-II is the evaluation of the fit of simulated perceived error curves to the sample-based (time) mean of perceived error measurements at all lead times considered. A fitting error say 95% of the time smaller than SEM at the 95% significance level indicates that experimental measurement data are consistent with the SAFE assumptions and error model. Fig. 8 displays the difference between the absolute value of the fitting error of perceived error variance and the 1.96SEM confidence interval at the 1.5-day lead time for variables U, T, and GH, for both SAFE-I and SAFE-II as follows:

$$| f_i^2 - \widehat{f_i^2} | - 1.96\mathrm{SEM}_i \ . \qquad (19)$$

The results at other lead times are qualitatively similar. The fitting error is smaller than 1.96SEM for all vertical levels for both SAFE-I and SAFE-II, which indicates that both error models are consistent with the measurements at the 95% confidence level. The lower negative values for SAFE-II indicate that the two additional parameters (the variance and decay rate of decaying errors) introduced in the present study offer a more complete error evolution model, attested by a closer fit to the perceived error measurements.

2) Total error

519 Fig. 9 displays the growing (red circles), decaying (blue circles), and total (black
520 circles) analysis error variance for the three variables investigated: U, T, and GH.
521 SAFE-I estimates of the total error variance (that is all assumed to be growing) are
522 shown as red crosses for a comparison. 6-hour lead time perceived error variance
523 measured as the difference between first guess and analysis fields is also provided
524 (green plus signs) as a possible indicator of analysis quality.

525 Looking first at the total error of the two model variables, U has an absolute
526 maximum around the upper-level jet (300 hPa), gradually/quickly dropping to
527 lower/much lower values near the bottom/top of the domain. In contrast, T has two
528 peaks, one presumably associated with the low-level jet (around 925 hPa), and a second
529 one above the jet level (200 hPa). Interestingly, the ratio between the maximum and
530 minimum total error variance in the vertical is in the 4-5 range for the two variables U
531 and T. The vertical profile of GH is less pronounced, with an absolute and secondary
532 maximum at 300 hPa and the surface, respectively.

533 As found earlier for selected variables in an OSSE setting (Section 5.a.1), when
534 no decaying errors are detected, SAFE-I error estimates (red crosses in Fig. 9) match
535 the total error variance estimates of SAFE-II (black open circles). In the presence of
536 decaying errors, SAFE-I still provides growing error estimates similar to SAFE-II;
537 however, these estimates are lower than the total error since the decaying analysis errors
538 are not directly accounted for.

539 6-hour perceived error variance directly relates to the quality of background
540 forecasts (or first guesses) in DA, and indirectly reflects error variance in the analysis.

Profiles of 6-hour perceived error variance for the three variables correlate well (at 0.88 or higher values) with SAFE-II estimates of total analysis error variance profiles (Fig. 9). As found in Section 5.a.2 for variables at 200 hPa height, perceived error measurements, however, are by a factor of 3-4 lower than estimates of true error through the entire profile of all variables. Such an underestimation can have profound impacts in the areas of data assimilation (underestimation of first guess error variance), ensemble initialization (specification of too low initial ensemble spread), and OSSE system calibration (setting simulated analysis error variance at too low levels).

We mention that the estimated U and T total error variance in Fig. 9 display similar vertical profiles to those measured directly by Privé et al. (2013a, see the thick solid lines in their Fig. 5a and d) and Privé and Errico (2013b, see the heavy dashed lines in their Fig. 1a and d) in their OSSE studies. The actual error variance values from their studies, however, differ from the operational forecast system error estimates in Fig. 9, just as was the case with the NCEP OSSE results (see related discussion in Section 5.a.3). Note that error levels may also differ due to distinctly different circulation regimes over the evaluation period of the operational vs OSSE DA-forecast system.

3) Growth rate

Beyond the variance and correlation of errors, SAFE-I and SAFE-II also provide estimates for the time evolution of error variance as a function of lead time. 6-hour amplification factors for (a) U, (b) T, and (c) GH are displayed as a function of height in Fig. 10. At all levels, GH has consistently faster error growth rate than the other two variables. For all variables, error growth peaks near the level of the mid-latitude jet

563 characterized with strong baroclinic instabilities at 300 hPa for U and GH, while around
564 450 hPa for T. Variations in growth rate across levels and variables reflect the instability
565 properties of different dynamical processes, operating on various spatial scales. The
566 slow error growth near the model top for the variables relative to other levels may be
567 explained by the strongly diffusive model dynamics (Houtekamer et al. 2005).

568 The model variables U and T also have a weaker maximum, near the low-level jet
569 and surface, respectively. It is interesting to point out that for the two model variables,
570 total analysis error variance has a corresponding maximum (typically 50 hPa for U and
571 150 hPa for T) above the double maxima observed in error growth. With vertically
572 uniform observational coverage, analysis error maxima are expected to exactly colocate
573 that of error growth. Given the density of in situ observations gradually decreases with
574 altitude, the upward shift of analysis maxima from growth rate maxima is expected.

575 *4) Decaying errors*

576 Just as shown for 200 hPa variables (Fig. 7 and Table 4), the decaying errors are
577 most prominent in GH fields throughout the entire atmosphere (Fig. 9). This is even
578 more visible in Fig. 11 that depicts the vertical profile of the percentage of the decaying
579 component in total analysis error variance (open circles) estimated by SAFE-II. As
580 discussed in Section 5.a.3, decaying errors in GH may be accentuated by the formula
581 used in their derivation from analyzed model prognostic variables.

582 In contrast, variable T is least affected by decaying errors, where they constitute
583 less than 15% of the total analysis error, and only in the upper half of the atmosphere
584 (see Figs. 9 and 11). U is in between, with two maxima situated around the upper and

585 lower level jets, with a spike near the top of the model. The source of decaying errors
586 in the analyzed variables includes representativeness errors (especially near diverse
587 topography, Quintana-Seguí et al., 2008; Jiménez and Dudhia, 2012), approximations
588 in balance constraints, observational noise, interpolations, localization, and other
589 statistical and numerical procedures in DA. Interestingly, no decaying errors are found
590 in the non-divergent mid-troposphere where commonly used balance constraints in the
591 DA schemes may be most applicable. As noted in Section 5.a.1, due to their nature,
592 decaying errors affect only analyses and short-range forecasts, therefore their estimates
593 are subject to a higher level of uncertainty. Further studies into the estimation of
594 decaying errors are therefore warranted.

595 *5) The decaying component of analysis increments*

596 An analysis field (Kalnay, 2003) is the sum of a first guess forecast that as we saw
597 itself contains decaying errors, and the analysis increment (AI) which is identical to the
598 6-hour percieved forecast error. It is well understood that the introduction of excessive
599 noise into the analysis via the AI in a cycled DA system can negatively affect the quality
600 of the analysis (Houtekamer and Mitchell, 2001; Dee, 2005). Hence the reduction of
601 noise in AI has been a prominent but hard to achieve goal in DA. As the noise introduced
602 by the data assimilation step via the AI into the analysis field contributes to the overall
603 level of decaying errors in the analysis, the vertical profile of the proportion of decaying
604 errors in the analysis is expected to be qualitatively similar to that in the AI.

605 Since both measurements and simulated values of 6-hour perceived forecast errors
606 are an integral part of SAFE-II, a convenient methodology offers itself for the

607 estimation of noise in the form of decaying errors in AI. The method is based on the
608 simulation of the variance in 6-hour LFDs with Eq. (10), and then fitting the simulated
609 curves to the sample mean of different lead time LFD variance measurements. A cost
610 function similar to Eq. (17) is used where x_i^2 and ω_i^{-1} are replaced by the sample mean
611 of 6-hour LFD variance and its SEM-based weight, respectively.

612 Fig. 11 shows the proportion of the decaying error component in the variance in AI
613 and analysis fields as crosses and open circles, respectively. As expected, the vertical
614 profile of the proportion of the decaying component of the analysis error is similar to,
615 though 10-30% higher than the decaying error component in AI for all variables
616 investigated. Just like in the analysis fields, decaying errors in AI are more pronounced
617 in the upper and lower parts of the model domain, roughly as those in analysis errors.
618 Note, however, that no decaying analysis errors are diagnosed for low-level temperature,
619 despite their presence in AI.

620 # 6. Conclusions

621 The evaluation of and improvements to data assimilation, ensemble forecasting,
622 and observing system simulation techniques require knowledge of error variance in
623 NWP analysis and forecast fields. Since reality is unknown, such error variance (i.e.,
624 "true" error variance) is directly not measurable. As observations are sporadic, most
625 systematic studies resort to estimating error variance by comparing forecast fields with
626 verifying analysis fields (i.e., "perceived" error). Such an approach (i) cannot assess
627 errors in the analysis, and (ii) ignores the effect of analysis error on forecast error
628 estimates.

629 Peña and Toth (2014) proposed an inverse procedure called Statistical Analysis
630 and Forecast Error (SAFE-I) algorithm for the bias-free estimation of true analysis and
631 forecast error variance. SAFE-I uses perceived error measurements (defined with
632 respect to the verifying analysis), and models them with a few parameters describing
633 the evolution of the true error in time: the initial error variance (g_0^2), the dynamical
634 growth rate (α), and the correlation between analysis and background forecast errors
635 (ρ_1). The unknown parameters are estimated by minimizing the difference between the
636 measured and modeled (via the unknown parameters) perceived error at various lead
637 times. SAFE-I is independent of assumptions and methods used in observing, DA, or
638 prediction systems.

639 An important assumption in SAFE-I is that at short lead times the true forecast
640 error variance (variances between forecasts and reality at the same time) grows
641 exponentially. This assumption, however, neglects any noise that observations or the
642 analysis procedure may inject into the analysis. Such errors typically project onto the
643 stable manifold of the system and thus rapidly decay, manifesting as a transitional
644 behavior in the evolution of the total error variance. In this paper, we relax the error
645 evolution assumption in SAFE-I by the introduction of decaying, in addition to the
646 growing analysis errors. Specifically, the modified SAFE method (SAFE-II) models
647 true forecast error variance as the sum of an exponentially growing, and an orthogonal
648 decaying component, the latter of which described by its variance (d_0^2) and decay rate
649 (β). The estimation of the expanded set of parameters in SAFE-II is facilitated by the

650 inclusion of additional measurements in the form of variances between lagged forecasts
651 valid at the same time, linked with parameters g_0^2 and α.

652 When decaying errors are present, the true forecast error variance may display an
653 initial transitional behavior, during which total error may decay or exhibit slower than
654 exponential growth while decaying errors diminish. Only after most decaying errors
655 vanish, does the total error assume an exponential pattern of growth.

656 The performance of SAFE-II was evaluated using the NCEP GFS/GSI system.
657 First, the assumptions behind SAFE-II were validated in an OSSE environment where
658 reality is exactly known. Area mean (Northern Hemisphere extratropics) true analysis
659 and forecast error variance was simulated by the error growth equation used in SAFE-
660 II, and fitted to sample-based measurements of these quantities from an OSSE system.
661 Error variance simulated by SAFE-II was found to be within sampling uncertainty of
662 the sample-based measurements. This, along with other related results indicate that the
663 assumptions behind SAFE-II are consistent with the experimental data.

664 Next, in the same OSSE environment, we pretended that truth is unknown and
665 used only perceived error measurements and SAFE-I or SAFE-II to produce and
666 validate against truth true error variance estimates. In the presence of decaying errors
667 (variable U), all SAFE-II error parameter estimates were found to be more accurate than
668 those with SAFE-I. For the two model variables tested (500 hPa U and T), SAFE-II
669 estimates of total analysis error variance were within 1% of the actual measured values,
670 while growth rate and error correlation values were within 2% of their reference values.
671 Growing analysis error variance estimates deviated less than 5% from their reference

values. Decaying errors were found to diminish rapidly. Hence perceived error measurements are affected only at a few early lead times, leading to larger (up to 50%) uncertainty in decaying error variance and decay rate estimates.

In Section 5, SAFE-II was used to estimate the error variance in operational NCEP analyses and forecasts. U, T, and GH perceived error variances simulated by SAFE-II were found to be within the sampling uncertainty of their measurement-based counterparts, indicating that the assumptions behind SAFE-II are consistent with the NCEP operational data. The key findings of this part of the study are as follows:

- The growth rate for the model variables U and T peaks around the upper-level jet in the areas of strongest baroclinic instabilities, with an error variance doubling time of around 23 hours. A weaker maximum appears around the lower level jet. Error variance doubling time near the surface is around 32 hours. Forecasts for GH exhibit error growth faster than those for U and T at all levels.
- The maximum of total analysis error variance for U and GH are near the upper-level jet (250-300 hPa), consistent with the level of their fastest error growth rate. The maximum of total analysis error variance for T is near the low-level jet (~ 925 hPa). Interestingly, the total analysis error for the model variables U and T peaks just above the maxima in growth rate. This may be explained by a general decrease in the density of in situ observations with increasing altitude.
- Decaying errors constitute up to 40% and 15% of the total analysis error variance for wind and temperature variables in the upper (and for U, also in the lower) atmosphere, respectively. Decaying errors originate from observational noise, and

approximations in DA procedures (e.g., improper balance constraints caused by model related errors near the model top, and lack of proper specification of representativeness error in areas of complex topography). No decaying errors are observed in the non-divergent mid-tropospheric region. This may be related to the quasi-nondivergent nature of dynamics at these layers where balancing the analysis variables is simpler and more straightforward.

- GH has a higher (50-60%) proportion of decaying errors than the model variables. This may be due to the accumulation of independent noise from the model variables as GH is derived from them.

7. Discussion

Due to the limited number of short lead time perceived error measurements influenced by decaying errors, the uncertainty in decaying parameter estimates are much higher than that in the estimates of the other parameters. Efficient approaches to constrain the estimates of decaying parameters and assess the uncertainty in such estimates need to be pursued further. The power law relationship of the error decorrelation (i.e. *Assumption 3*) may also need to be refined as decaying errors may not exhibit the same exponential-like decorrelation behavior as the growing errors.

Possible future applications of SAFE-II may also include gridpoint-wise estimation of analysis error variance. Geographical localization of excessive noise in analysis fields (e.g., due to a lack of physical or dynamical balance) may aid in the diagnosis and correction of weaknesses in DA techniques. Solid estimates of analysis uncertainty may also benefit ensemble initialization techniques.

A recurring observation in this study is that the commonly used perceived error variance gives a rather poor estimate of the true error variance (e.g., 3-4 fold underestimation at 6 hours lead time) and a related overestimation of the error growth rate within the first two days due to the neglect of (i) analysis errors and (ii) the correlation between error fields in the analyses and forecasts. The use of perceived error as an estimator of true error can have significant consequences in a number of areas:

Data Assimilation. In DA, background error variances will be underestimated. As DA performance depends only on the ratio (but not the absolute value) of errors in the background field vs. the observations, the tuning of DA schemes may lead to an underestimation of observational (including representativeness) errors as well.

Observing System Simulation Experiments. If true error variance is assumed to be as low as perceived error variance measured in operational forecast systems, OSSE systems may be tuned to exhibit too low true error variance. This problem may be evidenced in NOAA's OSSE system (cf. column 7 in Tables 2 and 4).

Predictability. When the growth of perturbations such as lagged forecast differences (LFD) is compared with the growth of perceived error, the latter, since at short lead times perceived errors have unrealistically low values, appears to be significantly faster than the former. This situation has been widely interpreted in the literature as a sign that external predictability is shorter than internal predictability (i.e., the divergence of trajectories in nature is faster than in its numerical models, e.g. Simmons et al., 1995). True error, however, amplifies much slower than perceived error (see, e.g., Figs. 3 and 7), possibly eliminating the need for such hypothetical

738 explanations. Implications may include a longer than currently thought limit on
739 predictability.

740 *Ensemble Forecasting*. If the size of initial perturbations is set so that 6-hour
741 ensemble variance matches 6-hour perceived error variance, the ensemble, though it
742 may appear reasonable when its spread is checked against perceived error, actually will
743 start out underdispersive. Irrespective of initial perturbation generation methods, the
744 underdispersiveness readily manifests itself at later lead times (e.g., Buizza et al. 2005),
745 however, when analysis error variance becomes negligible compared to forecast error
746 variance. Conventionally, the situation is explained as insufficient perturbation growth
747 due to model imperfectness presumedly related to the numerical models being more
748 predictable than the atmosphere (i.e., too high internal predictability). The notion and
749 an array of stochastic model perturbation methods (Buizza et al., 1999; Shutts, 2005)
750 have been proposed to hasten perturbation growth with the aim of remedying a problem
751 that may not exist. Future studies can further explore the validity of the Statistical
752 Analysis and Forecast Error (SAFE) estimation based interpretations advanced above.

753

754 ***Acknowledgments.*** Drs. Lidia Cucurull, Ruifang Li, and Tanya Peevey kindly provided
755 the Observational System Simulation Experiment data and the corresponding
756 references. Discussions with Drs. Krishna Kumar (NCEP), Jordan Alpert (NCEP),
757 Fanglin Yang (NCEP), Si Shen (NCAR), and Prof. Roman Krzysztofowicz (University
758 of Virginia) are gratefully acknowledged. We acknowledge the encouragement and
759 support of Kevin Kelleher, former Director of GSD.

760 Appendix A

761 **Sampling Uncertainty**

762 Just as SAFE-I (Peña and Toth 2014), SAFE-II estimates the unknown parameters
763 of true analysis and forecast error variance by fitting perceived error variance modeled
764 with the unknown parameters to sample-based measurements of perceived error
765 variance. The expected error in finite sample-based estimates of the expected value of
766 normally distributed variables is given by the Standard Error of the Mean (or
767 Measurement, SEM):

768 $$\mathrm{SEM}_i = sd_i \cdot f / \sqrt{N}, \quad \text{(A1)}$$

769 where sd_i represents the sample standard deviation in the sample at lead time i, N is the
770 sample size, and $f = \sqrt{(1+r_1)(1-r_1)^{-1}}$ is an adjustment coefficient accounting for serial
771 correlation (r_1) in the sample.

772 As the standard deviation of finite sample-based mean tend to grow with lead time,
773 observed quantities at longer time ranges will need to be given smaller weight in the
774 minimization procedure. The standardized weights w_i in Eq. 8 are defined as:

775 $$w_i = \mathrm{SEM}_i / \sum_i \mathrm{SEM}_i. \quad \text{(A2)}$$

776 Note that the definition of SEM_i and w_i can be generalized to other finite sample-based
777 estimates of expected value, like the lagged forecast difference and true forecast error
778 variance et al. used in this study.

779 Since SEM values quantify the uncertainty in sample mean values, they can also
780 be considered as confidence intervals when SAFE estimates are compared with the
781 mean of measurements. Assuming that the finite-sample mean of perceived error
782 variance follows a Gaussian distribution, the 95% confidence interval can be defined

783 by adding and subtracting 1.96 times the SEM_i value to/from the perceived error
784 variance measurements.

785 Appendix B

786 **OSSE Setup**

787 In OSSEs, analyses and forecasts are generated the same way as in an operational
788 NWP system, except the role of real observations are taken by simulated observations.
789 A long integration with a fine resolution model other than that used in the NWP DA-
790 forecast system is usually considered as truth (or nature), from which simulated
791 observations are generated with the addition of noise meant to represent different
792 sources of observational and representativeness errors (e.g., Atlas, 1997). Since truth is
793 exactly known, when carefully designed, OSSEs offer a unique and fully controlled
794 environment in which to evaluate the quality of NWP techniques.

795 Nature used in this OSSE system was created by the European Center for Medium-
796 Range Weather Forecasts (ECMWF) operational model version c31r1 at T511 (about
797 40 km) horizontal and 91-level vertical resolution, with boundary forcing data from 1
798 May 2005 to 31 May 2006 (Masutani et al., 2006; Andersson and Matsuani, 2010). The
799 NWP modeling (GFS) and DA system (Gridpoint Statistical Interpolation analysis -
800 GSI) are based on an earlier and reduced resolution (T382, about 52 km, and 64-level)
801 version of NCEP's operational suite with a non-hybrid DA scheme. The observations
802 assimilated include conventional, satellite, and COSMIC-2 ("Constellation for
803 Observing System for Meteorology, Ionosphere, and Climate") data generated from the
804 nature run. Representativeness errors are inherent in the simulated observations due to
805 a difference in resolution between the nature run and the NWP system. No systematic

or random errors were otherwise added to nature for the simulated observations, except for satellite radiances. All observations are assimilated using a ±1 hour window centered at nominal analysis times.

REFERENCES

Andersson, E., and Matsutani, M. (2010) Collaboration on observing system simulation experiments (joint OSSE). *ECMWF Newsletter*, **123**, 14–16.

Atlas, R. (1997) Atmospheric observations and experiments to assess their usefulness in data assimilation. *J. Meteor. Soc. Japan*, **75**, 1–20.

Buizza, R., Miller M., and Palmer T. N. (1999) Stochastic representation of model uncertainties in the ECMWF Ensemble Prediction System. *Quart. J. Roy. Meteor. Soc.*, **125**, 2887–2908.

Buizza, R., Houtekamer, P. L., Toth, Z., Pellerin, G., Wei, M. and co-authors (2005) A comparison of the ECMWF, MSC and NCEP global ensemble prediction systems. *Mon. Wea. Rev.*, **133**, 1067–1097.

Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C. (1995) A limited memory algorithm for bound constrained optimization. *SIAM J. Sci. Stat. Comput.*, **16(5)**, 1190–1208.

Cucurull, L., Li, R. and Peevey, T. R. (2017) Assessment of radio Occultation Observations from the COSMIC-2 Mission with a Simplified Observing System Simulation Experiment Configuration. *Mon. Wea. Rev.*, **145**, 3581–3597.

Dalcher, A., Kalnay, E. , and Hoffman, R. N. (1988) Medium range lagged average forecasts. *Mon. Wea. Rev.*, **116**, 402–416.

Dee, D. (2005) Bias and data assimilation. *Quart. J. Roy. Meteor. Soc.*, **131,** 3323–3343.

Ding, R. Q., and Li, J. P. (2007) Nonlinear finite-time Lyapunov exponent and predictability. *Phys. Lett.*, **364A**, 396–400.

833 Ding, R. Q., Li, J., and Li, B. S. (2017) Determining the Spectrum of the Nonlinear Local
834 Lyapunov Exponents in a Multidimensional Chaotic System. *Adv. in Atmos. Sci.*, **34(9)**,
835 1027–1034.
836 Feng, J., Ding, R. Q., Liu, D. Q. and Li, J. P. (2014) The application of nonlinear local
837 Lyapunov vectors to ensemble predictions in the Lorenz systems. *J. Atmos. Sci.*, **71(9)**,
838 3554–3567.
839 Feng, J., Toth, Z., and Peña, M. (2017) Spatially extended estimates of analysis and short-range
840 forecast error variances. *Tellus*, **69A**, 1325301.
841 Feng, J., Li, J., Ding, R. and Toth, Z. (2018) Comparison of nonlinear local Lyapunov vectors
842 and bred vectors in estimating the spatial distribution of error growth. *Journal of the*
843 *Atmospheric Sciences*, **75**(4), 1073–1087.
844 Fisher, M. (1996) The specification of background error variances in the ECMWF variational
845 analysis system. In: *Proceedings ECMWF Seminar on Data Assimilation*. Reading, UK,
846 Available from ECMWF, pp. 645–652.
847 Gilmour, I., Smith, L. A., and Buizza, R. (2001) Linear regime duration: Is 24 hours a long time
848 in synoptic weather forecasting? *J. Atmos. Sci.*, **58,** 3525–3539.
849 Grell, G. A., Dudhia, J., and Stauffer, D. (1995) A description of the fifth-generation PENN
850 State/NCAR Mesoscale Model (MM5). NCAR Tech. Note NCAR/TN-398+STR, 10 pp.
851 Hamill, T., Snyder, M., C., and Morss, R. E., (2002) Analysis-error statistics of a
852 quasigeostrophic model using three-dimensional variational assimilation. *Mon. Wea.*
853 *Rev*, **130,** 2777–2791.
854 Hamill, T. M., and Whitaker, J. S. (2011) What constrains spread growth in forecasts initialized
855 from ensemble Kalman filters? *Mon. Wea. Rev.*, **139**, 117–131.
856 Houtekamer, P. L., and Mitchell, H. L. (2001) A sequential ensemble Kalman filter for
857 atmospheric data assimilation. *Mon. Wea. Rev*, **129,** 123–137.
858 Houtekamer, P. L., Mitchell, H. L., Pellerin, G., Buehner, M., Charron, M., Spacek, L.,
859 and Hansen, B. (2005) Atmospheric data assimilation with an ensemble Kalman filter:
860 Results with real observations. *Mon. Wea. Rev.*, **133,** 604–620.
861 Huang, X-Y., and Lynch, P. (1993) Diabatic digital-filtering initialization: Application to the
862 HIRLAM model. *Mon. Wea. Rev.*, **121,** 589–603.

Hunt, B., Kostelich, E., and Szunyogh, I. (2007) Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter. *Physica D*, **230,**112–126.

Jiménez, P. A., and Dudhia, J. (2012) Improving the representation of resolved and unresolved topographic effects on surface wind in the WRF model. *J. Appl. Meteor. Climatol.*, **51**, 300–316.

Kalnay, E. (2003) Atmospheric Modeling, Data Assimilation and Predictability. Cambridge University Press, 341 pp.

Kleist, D. T., Parrish, D. F., Derber, J. C., Treadon, R., Errico, R. M., and Yang, R. (2009) Improving incremental balance in the GSI 3DVAR analysis system. *Mon. Wea. Rev.*, **137,** 1046–1060.

Legras, B., and Vautard, R. (1996) A guide to Lyapunov vectors. Proc. ECMWF Seminar on Predictability, Vol. 1, Reading, United Kimgdom, ECMWF, 143–156. [Available from ECMWF, Shinfield Park, Reading, Berkshire RG2 9AX, United Kingdom.]

Li, J. P., and Ding, R. Q. (2011) Temporal-spatial distribution of atmospheric predictability limit by local dynamical analogues. *Mon. Wea. Rev.,* **139**, 3265–3283.

Lorenc, A. C. (2003) The potential of the ensemble kalman filter for NWP – a comparison with 4D-Var. *Q. J. R. Meteorol. Soc.*, **129**, 3183–3203.

Lorenz, E. N. (1963) Deterministic nonperiodic flow. *J. Atmos. Sci.* **20**, 130141.

Lorenz, E. N. (1982) Atmospheric predictability experiments with a large numerical model. *Tellus 34*, 505–513.

Lorenz, E. N. (1996) Predictability: A problem partly solved. *Proc. ECMWF Seminar on Predictability,* Vol. I, Reading, United Kingdom, ECMWF, 1–18.

Masutani, M., Woollen, J., Lord, S., Kleespies, T., Emmitt, G. and Co-authors (2006) Observing System Simulation Experiments at NCEP. Office Note 451. National Centers for Environmental Prediction. College Park, Maryland, USA.

Mitchell, H. L., Houtekamer, P. L., and Pellerin, G. (2002) Ensemble size, balance, and model-error representation in an ensemble Kalman filter. *Mon. Wea. Rev.*, **130,** 2791–2808.

Molteni, F., Buizza R., Palmer, T. N., and Petroliagis, T. (1996) The new ECMWF Ensemble Prediction System: Methodology and validation. *Quart. J. Roy. Meteor. Soc.*, **122**, 73–119.

892 Nicolis, C., Perdigao, R. and Vannitsem, S. (2009) Dynamics of prediction errors under the
893 combined effect of initial condition and model errors. *J. Atmos. Sci.*, **66**, 766–778.
894 Orrell, D., Smith, L. , Barkmeijer, J. , and Palmer, T. N. (2001) Model error in weather
895 forecasting. *Nonlinear Processes Geophys.*, **8,** 357–371.
896 Palatella, L., Carrassi, A., and Trevisan, A. (2013) Lyapunov vectors and assimilation in the
897 unstable subspace: Theory and applications. *J. Phys. A: Math. Theor.*, **46**, 254020.
898 Peña, M., and Toth, Z. (2014) Estimation of analysis and forecast error variances. *Tellus*, **66A**,
899 21767.
900 Peña, M., Toth, Z., and Wei, M. (2010) Controlling noise in ensemble data assimilation
901 schemes. *Mon. Wea. Rev.*, **138**, 1502–1512.
902 Pires, C., Vautard, R. , and Talagrand, O. (1996) On extending the limits of variational
903 assimilation in nonlinear chaotic systems. *Tellus*, **48A,** 96–121.
904 Privé, N. C., Errico, R. M., and Tai, K.-S. (2013a) The influence of observation errors on
905 analysis error and forecast skill investigated with an observing system simulation
906 experiment. *J. Geophys. Res. Atmos.*, **118**, 5332–5346.
907 Privé, N., and Errico, R. M. (2013b) The role of model and initial condition error in numerical
908 weather forecasting investigated with an observing system simulation experiment. *Tellus-*
909 *A*, **65**, 21740.
910 Quintana-Seguí, P., and Coauthors (2008) Analysis of near-surface atmospheric variables:
911 Validation of the SAFRAN analysis over France. *J. Appl. Meteor. Climatol.*, **47,** 92–107.
912 Shutts, G. J. (2005) A kinetic energy backscatter algorithm for use in ensemble prediction
913 systems. *Quart. J. Roy. Meteor. Soc.*, **131**, 3079–3102.
914 Simmons, A. J., Mureau, R., and Petroliagis, T. (1995) Error growth estimates of predictability
915 from the ECMWF forecasting system. *Quart. J. Roy. Meteor. Soc.,* **121,** 1739–1771.
916 Stewart, L. M., Dance, S. L., and Nichols, N. K. (2013) Data assimilation with correlated
917 observation errors: Experiments with a 1-D shallow water model. *Tellus*, **65A**, 19546.
918 Snyder, C., and Hamill, T. M. (2003) Leading Lyapunov vectors of a turbulent baroclinic jet in
919 a quasigeostrophic model. *J. Atmos. Sci.*, **60,** 683–688.
920 Toth, Z., and Kalnay, E. (1993) Ensemble forecasting at NMC: the generation of perturbations.
921 *Bull. Amer. Meteor. Soc.*,**74**, 2317–2330.

922 Toth, Z., and Kalnay, E. (1997) Ensemble Forecasting at NCEP: the breeding method. *Mon.*
923 *Wea. Rev.*, **125**, 3297–3318.
924 Trevisan, A., and Legnani, R. (1995) Transient error growth and local predictability: a study in
925 the Lorenz system. *Tellus*, **47A**, 103–117.
926 Trevisan, A., and Uboldi, F. (2004) Assimilation of standard and targeted observations within
927 the unstable subspace of the observation–analysis–forecast cycle system. *J. Atmos. Sci.*,
928 **61**, 103–113.
929 Vannitsem, S., and Nicolis, C. (1994) Predictability experiments on a simplified thermal
930 convection model: The role of spatial scales. *J. Geophys. Res.*, **99,** 10377–10385.
931 Vannitsem, S., and Nicolis, C. (1997) Lyapunov vectors and error growth patterns in a T21L3
932 quasigeostrophic model. *J. Atmos. Sci.,***54,** 347–361.
933 Vannitsem, S., and Toth, Z. (2002) Short-term dynamics of model errors. *J. Atmos. Sci.*, **59**,
934 2594–2604.
935 Wee, T.-K., Kuo Y.-H. , Lee D.-K. , Liu Z., Wang W., and Chen S.-Y. (2012) Two overlooked
936 biases of the Advanced Research WRF (ARW) model in geopotential height and
937 temperature. *Mon. Wea. Rev.*, **140**, 3907–3918.
938 Wei, M., Toth, Z., Wobus, R. , and Zhu, Y. (2008) Initial perturbations based on the ensemble
939 transform (ET) technique in the NCEP global operational forecast system.*Tellus*, **60A,** 62–
940 79.
941 Whitaker, J. S., Hamill, T. M. , Wei, X., Song, Y. , and Toth, Z. (2008) Ensemble data
942 assimilation with the NCEP Global Forecast System. *Mon. Wea. Rev.*, **136,** 463–482.
943 Wolf, A., Swift, J. B., Swinney, H. L., and Vastano, J. A. (1985) Determining Lyapunov
944 exponents from a time series. *Physica D*, **16**, 285–317.
945 Yang, F. (2016) Evaluation of hurricane forecast skills of NCEP GFS retrospective experiments
946 for the FY2016 implementation. *32nd Conf. on Hurricanes and Tropical Meteorology.*
947 *2016*, San Juan, Puerto Rico.
948 https://ams.confex.com/ams/32Hurr/webprogram/Paper293991.html.
949 Ziehmann, C., Smith, L. A., and Kurths, J. (2000) Localized Lyapunov exponents and the
950 prediction of predictability. *Phys. Lett*, **271A,** 237–251.
951

Figure Captions

TABLE 1. Summary of Assumptions 1-6 behind the SAFE-II method related to the use of perceived error variance (PEV) and variance of lagged forecast difference (VLFD) measurements.

TABLE 2. Comparison between fitted (Fit) and reference (Ref) values of error parameters for zonal wind (U), temperature (T), and geopotential height (GH) at 500 hPa in the Observing System Simulation Experiments (OSSE) using SAFE-I and SAFE-II, respectively. g_0^2 and d_0^2 denote the growing and decaying components, with Δt=6 hours growth and decay rates of $e^{\Delta t \cdot \alpha}$ and $e^{\Delta t \cdot \beta}$. The values in brackets indicate the percentage of estimation error compared to ground truth (reference). Entries with – indicates where parameter values are not available. The rightmost column lists the growth rate of lagged forecast difference (LFD) variance per 6 hours.

TABLE 3. SAFE-I and SAFE-II estimates of error evolution parameters for 500 hPa U and T in the OSSE experiments. Reference values (Ref) are the SAFE-II fitted parameter values from Table 2. The values in brackets indicate the 95% sampling uncertainty confidence intervals of Ref.

TABLE 4. Estimated error parameters for U, T, and GH at 200 and 500 hPa in GFS operational forecasts using SAFE-II.

FIG 1. Schematic of the evolution of the true forecast error variance (solid) and its growing (dashed) and decaying (dotted) components.

FIG 2. 3D schematic of the relationship between the correlations of true ($\mathbf{TF}_{i\text{-}1}$ and $\mathbf{TF}_i$, $\rho_{i\text{-}1,i}$) and perceived errors ($\mathbf{AF}_{i\text{-}1}$ and $\mathbf{AF}_i$, $\gamma_{i\text{-}1,i}$), all valid at the same time. **F**, **A**, and **T** represent forecast, analyzed, and true states, respectively.

FIG 3. Sample-mean based estimates of ground truth for true forecast error variance (open circles with 95% vertical confidence intervals as vertical bars) along with the corresponding fitted values (solid line) for variables (a) U, (b) T, and (c) GH at 500 hPa in the OSSE environment. For comparison, perceived error variance measurements are also shown as dashed lines.

FIG 4. Estimates of growing (dashed line), decaying (dotted line) and total (solid line) error variance along with the corresponding fitted values (hollow circles with 95% confidence intervals as vertical bars) for GH at 500 hPa in the OSSE environment.

FIG 5. Sample-mean fitted (solid line) and measured (circle) ρi for variables (a) U, (b) T and (c) GH at 500 hPa in the OSSE. Vertical bars represent 95% confidence intervals.

FIG 6. Comparison between the correlations of lagged perceived (circle) and true forecast errors (cross) with a lag of 24 hours as a function of lead time for variables (a) U, (b) T, and (c) GH at 500 hPa in the OSSE.

FIG 7. Temporal variation of sample-mean actual (circle) and simulated (black thin line) perceived error variances and estimated total (black thick line), growing (red) and decaying (blue) true forecast error variances over Northern Hemisphere for variables (a) U, (b) T, and (c) GH at 200 hPa for GFS-GSI operational forecast system.

FIG 8. Profile of the differences between the absolute fitting errors and the 95% confidence interval of perceived error variances at 1.5 days by SAFE-II (circle) and SAFE-I (cross) for variables (a) U, (b) T and (c) GH for GFS-GSI operational forecast system.

FIG 9. Same as Figure 8, but for estimated total (black circle), growing (red circle) and decaying (blue circle) analysis error variance by SAFE-II. Green and Red crosses represent the 6-hour perceived error variance and the estimated analysis error variance by SAFE-I, respectively.

FIG 10. Same as Figure 8 but for the estimated growth rate of error variance per 6 hours.

FIG 11. Same as Figure 8 but for estimated percentage of decaying components in total analysis error variance by SAFE-II (circle) and variance of analysis increment (cross).

1008 TABLE 1. Summary of Assumptions 1-6 behind the SAFE-II method related to the

1009 use of perceived error variance (PEV) and variance of lagged forecast difference

1010 (VLFD) measurements.

Subject	Estimation area	Assumption	Section introduced / validated
1. Model error	PEV	Negligible for studied variables	2.a / 4.a
2.Error evolution	PEV	Exponential growth / decay of initial error variance	2.a, b / 4.a
3.Data impact on analysis	PEV	Power law decorrelation of analysis error from increasing lead time foecast error	2.a / 4.a
4.Relationship between true and perceived error variance	VLFD	True and perceived error variances become similar with longer lead times	2.c / 4.b
5.Transient period	VLFD	Decaying errors diminish in first 24 hours of integration	2.c / 4.a
6.Divergence rate of model trajectories	VLFD	Divergence rate is similar between lagged forecasts vs. forecast and truth	2.c / 4.a, b

1011

TABLE 2. Comparison between fitted (Fit) and reference (Ref) values of error parameters for zonal wind (U), temperature (T), and geopotential height (GH) at 500 hPa in the Observing System Simulation Experiments (OSSE) using SAFE-I and SAFE-II, respectively. g_0^2 and d_0^2 denote the growing and decaying components, with Δt=6 hours growth and decay rates of $e^{\Delta t \cdot \alpha}$ and $e^{\Delta t \cdot \beta}$. The values in brackets indicate the percentage of estimation error compared to ground truth (reference). Entries with – indicates where parameter values are not available. The rightmost column lists the growth rate of lagged forecast difference (LFD) variance per 6 hours. The units of the error variances (g_0^2 and d_0^2) for U, T, and GH are $(\text{m s}^{-1})^2$, K^2, and m^2, respectively.

		g_0^2	$e^{\Delta t \cdot \alpha}$	d_0^2	$e^{\Delta t \cdot \beta}$	$g_0^2+d_0^2$	$\frac{d_0^2}{g_0^2+d_0^2}$	LFD var growth
U :	**Fit : SAFE-I**	2.09	1.157	0.0	-	2.09 (8.0%)	0.0	
	Fit : SAFE-II	1.96	1.168	0.248	0.221	2.21 (2.6%)	11.2%	1.146
	Ref / 1.96SEM	-	-	-	-	2.27 / 0.265	-	
T :	**Fit : SAFE-I**	0.229	1.174	0.0	-	0.229 (9.0%)	0.0	
	Fit : SAFE-II	0.229	1.174	0.0	-	0.229 (9.0%)	0.0	1.169
	Ref / 1.96SEM	-	-	-	-	0.210 / 0.024	-	
GH :	**Fit :SAFE-I**	14.9	1.288	0.0	-	14.9 (12.9%)	0.0	
	Fit : SAFE-II	13.1	1.318	4.34	0.368	17.5 (2.2%)	24.8%	1.278
	Ref / 1.96SEM	-	-	-	-	17.1 / 2.10	-	

1025 TABLE 3. SAFE-I and SAFE-II estimates of error evolution parameters for 500 hPa U
1026 and T in the OSSE experiments. Reference values (Ref) are the SAFE-II fitted
1027 parameter values from Table 2. The values in brackets indicate the 95% sampling
1028 uncertainty confidence intervals of Ref.

		g_0^2	$e^{\Delta t \cdot \alpha}$	d_0^2	$e^{\Delta t \cdot \beta}$	x_0^2	$d_0^2 / (g_0^2+d_0^2)$	ρ_1
U :	**SAFE-I**	2.07	1.149	0.0	-	2.07	0.0	0.792
	SAFE-II	2.06	1.153	0.19	0.34	2.25	8.4%	0.804
	Ref	1.96	1.168	0.25	0.22	2.27 (0.27)	11.2%	0.796 (0.023)
T :	**SAFE-I**	0.21	1.165	0.0	-	0.21	0.0	0.810
	SAFE-II	0.21	1.165	0.0	-	0.21	0.0	0.810
	Ref	0.23	1.174	0.0	-	0.21 (0.024)	0.0	0.824 (0.031)

1029

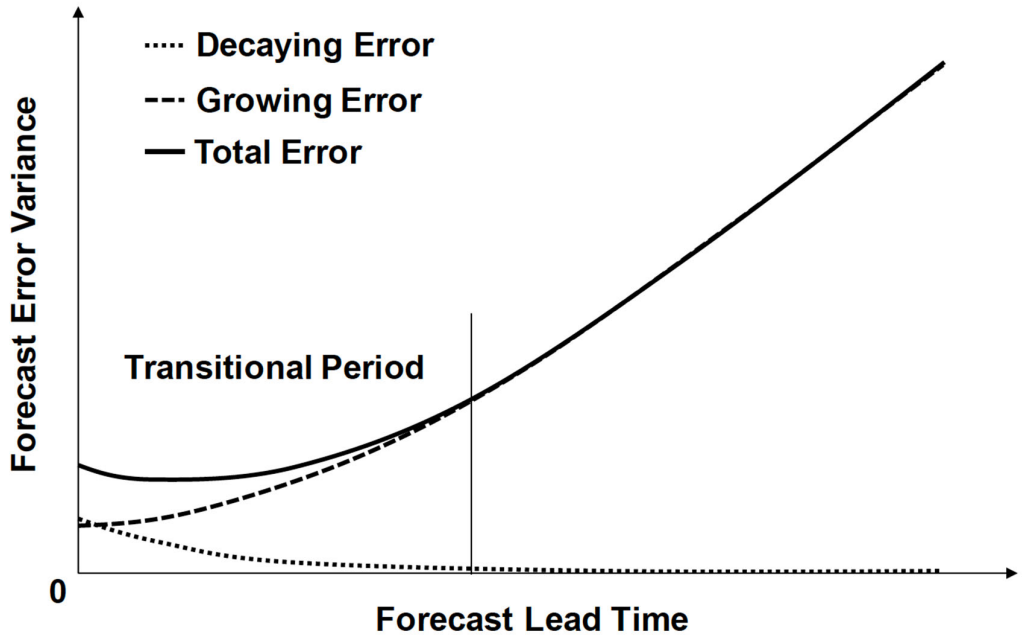
1030 TABLE 4. Estimated error parameters for U, T, and GH at 200 and 500 hPa in GFS

1031 operational forecasts using SAFE-II.

		g_0^2	$e^{\Delta t \cdot \alpha}$	d_0^2	$e^{\Delta t \cdot \beta}$	$g_0^2+d_0^2$	$d_0^2 / (g_0^2+d_0^2)$	ρ_1
U :	200 hPa	3.74	1.17	1.93	0.37	5.67	34.0%	0.87
	500 hPa	3.67	1.16	0.0	-	3.67	0.0	0.83
T :	200 hPa	0.39	1.19	0.049	0.35	0.439	11.2%	0.86
	500 hPa	0.25	1.21	0.0	-	0.25	0.0	0.84
GH :	200 hPa	34.60	1.32	39.47	0.14	74.07	53.3%	0.87
	500 hPa	24.72	1.32	34.88	0.14	59.60	58.5%	0.87

1032

1033



1034

1035 Figure 1. Schematic of the evolution of the true forecast error variance (solid) and its

1036 growing (dashed) and decaying (dotted) components.

1037

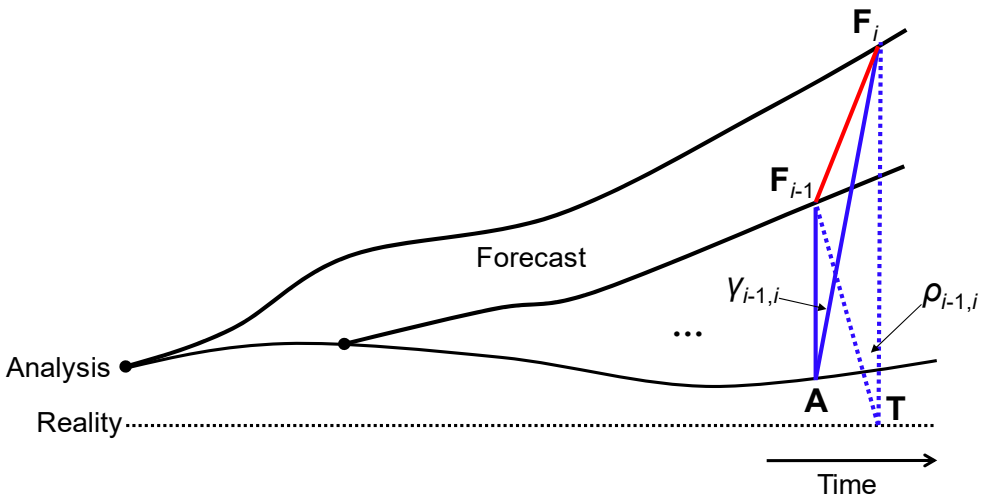
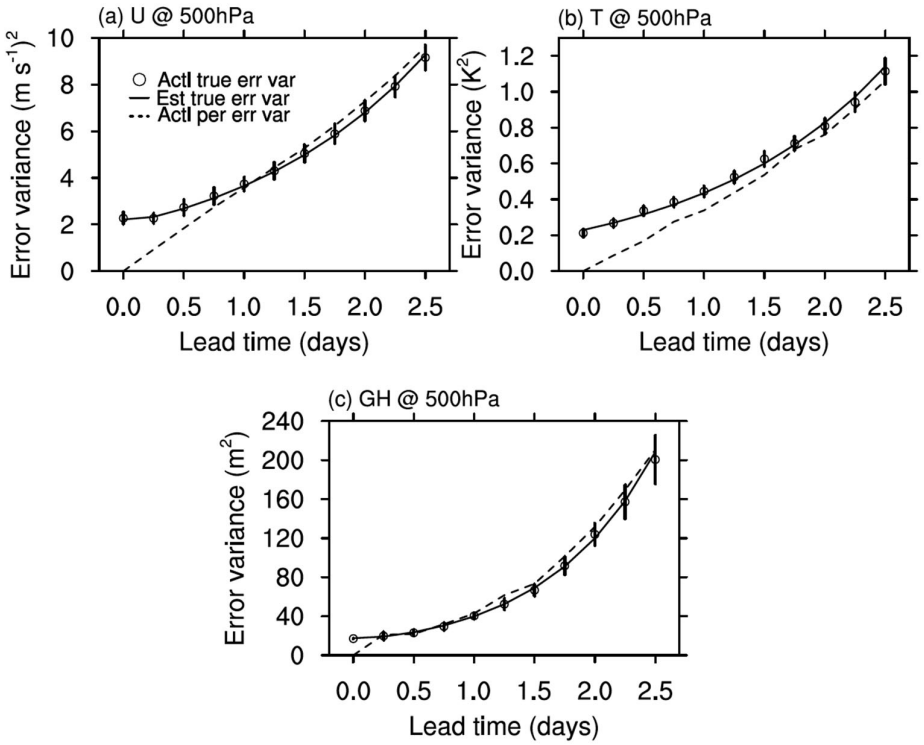


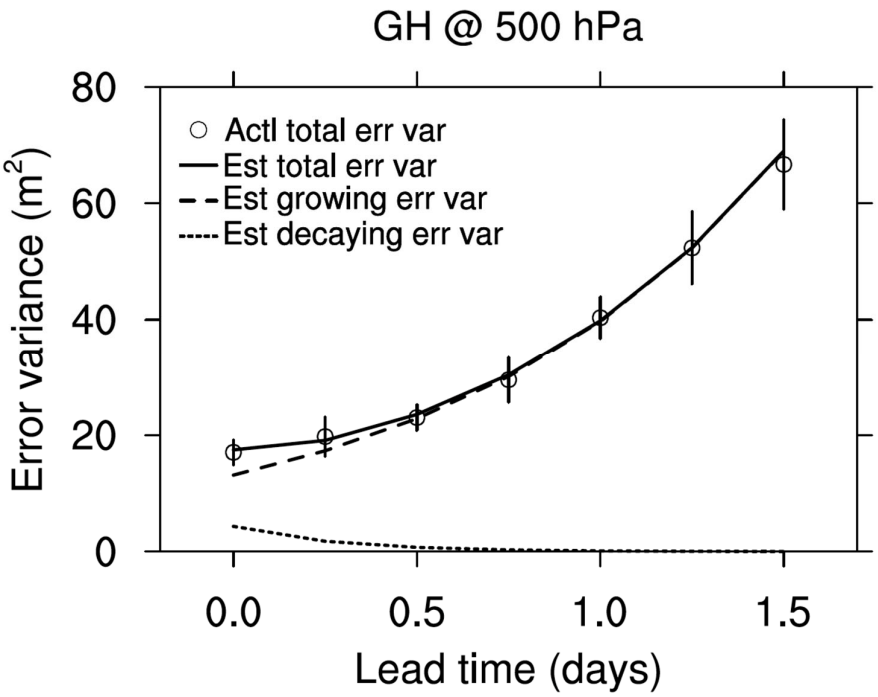
Figure 2. 3D schematic of the relationship between the correlations of true ($\mathbf{TF}_{i\text{-}1}$ and $\mathbf{TF}_i$, $\rho_{i\text{-}1,i}$) and perceived errors ($\mathbf{AF}_{i\text{-}1}$ and $\mathbf{AF}_i$, $\gamma_{i\text{-}1,i}$), all valid at the same time. **F**, **A**, and **T** represent forecast, analyzed, and true states, respectively.



1043

1044 Figure 3. Sample-mean based estimates of ground truth for true forecast error variance
1045 (open circles with 95% vertical confidence intervals as vertical bars) along with the
1046 corresponding fitted values (solid line) for variables (a) U, (b) T, and (c) GH at 500 hPa
1047 in the OSSE environment. For comparison, perceived error variance measurements are
1048 also shown as dashed lines.

1049



1050

1051 Figure 4. Estimates of growing (dashed line), decaying (dotted line) and total

1052 (solid line) error variance along with the corresponding fitted values (hollow

1053 circles with 95% confidence intervals as vertical bars) for GH at 500 hPa in the

1054 OSSE environment.

1055

1056

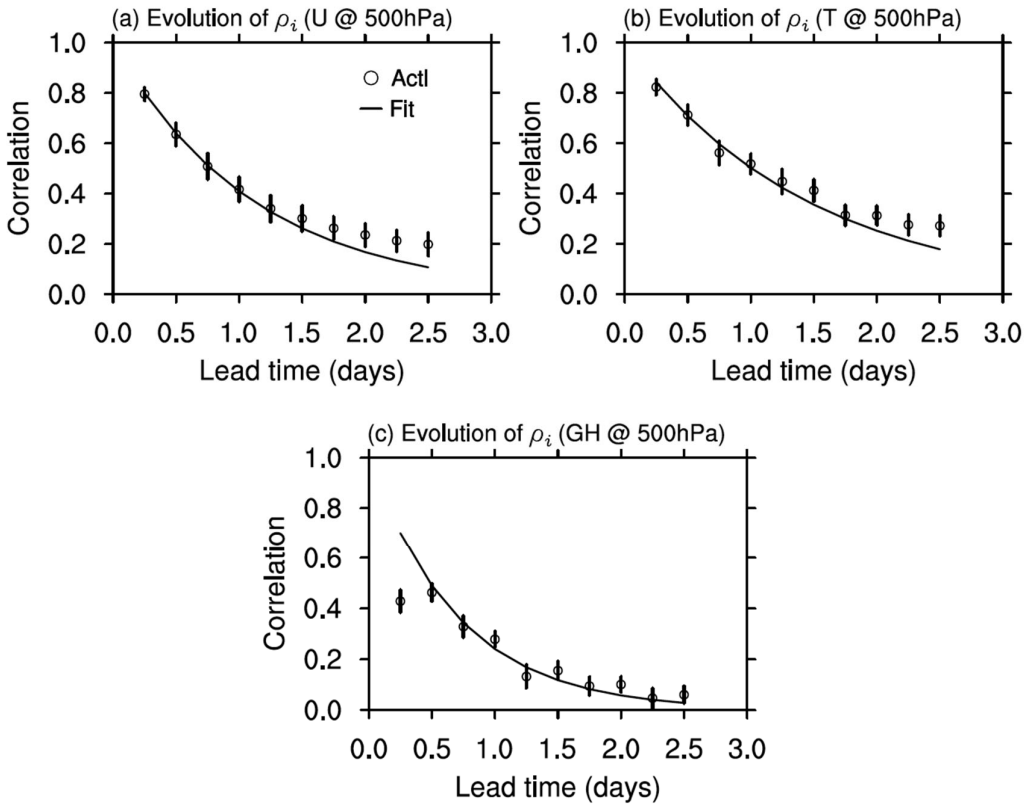
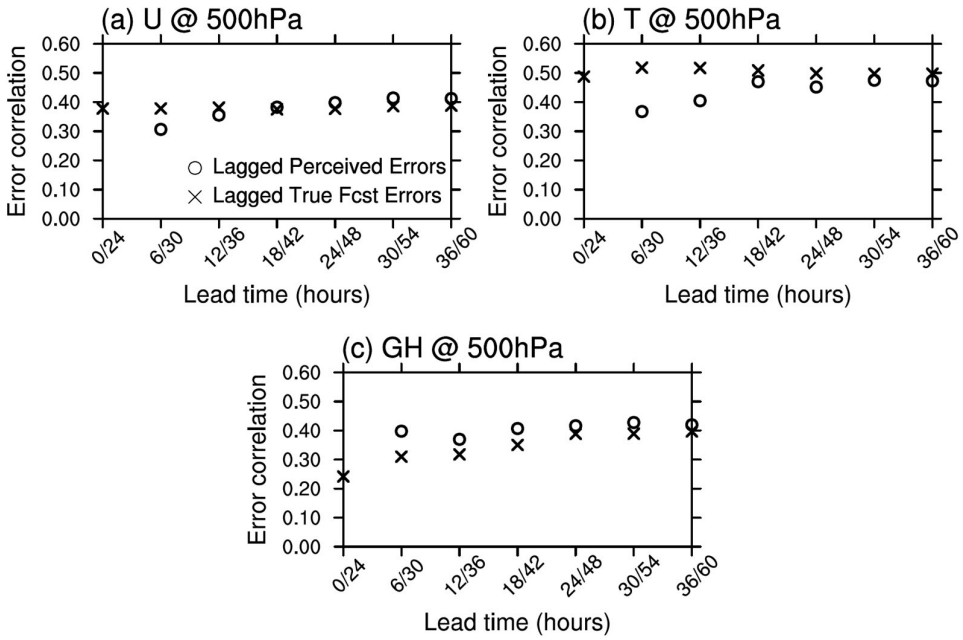


Figure 5. Sample-mean fitted (solid line) and measured (circle) ρ_i for variables (a) U, (b) T, and (c) GH at 500 hPa in the OSSE. Vertical bars represent 95% confidence intervals.

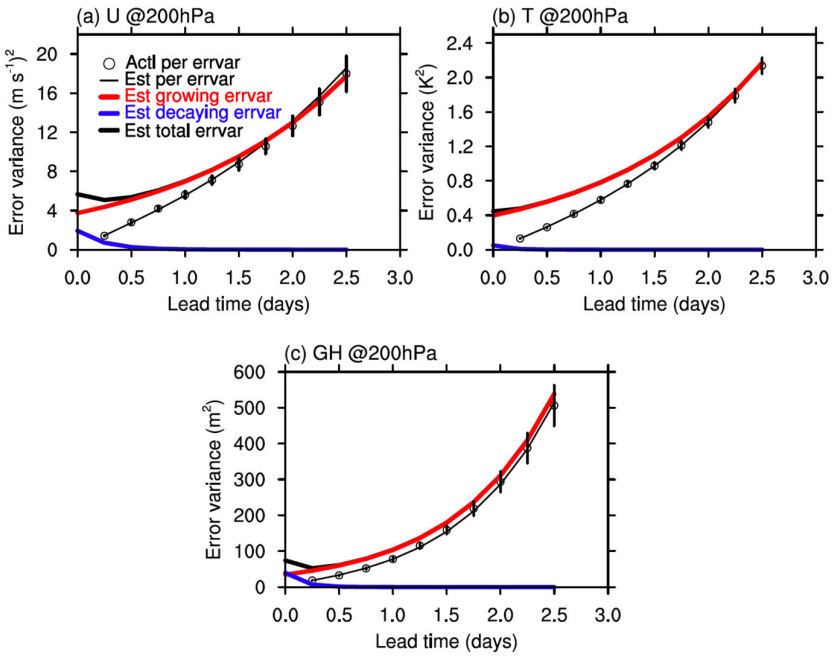


1063

1064 Figure 6. Comparison between the correlations of lagged perceived (circle) and

1065 true forecast errors (cross) with a lag of 24 hours as a function of lead time for

1066 variables (a) U, (b) T, and (c) GH at 500 hPa in the OSSE.



1067

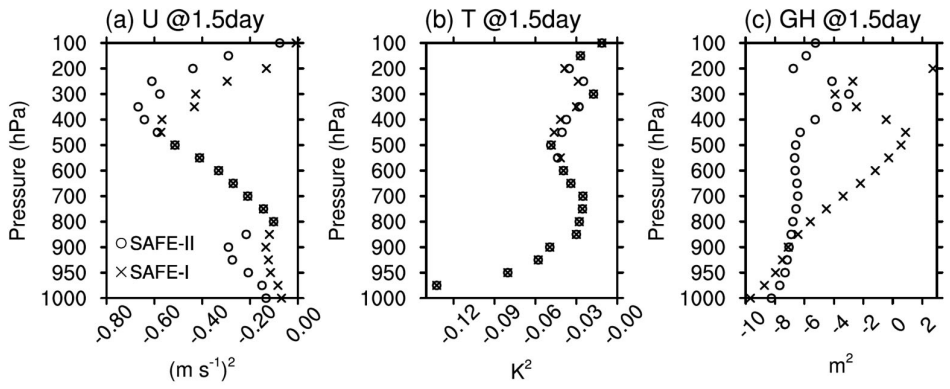
1068 Figure 7. Temporal variation of sample-mean actual (circle) and simulated (black thin

1069 line) perceived error variances and estimated total (black thick line), growing (red) and

1070 decaying (blue) true forecast error variances over Northern Hemisphere for variables

1071 (a) U, (b) T, and (c) GH at 200 hPa for GFS-GSI operational forecast system.

1072



1073

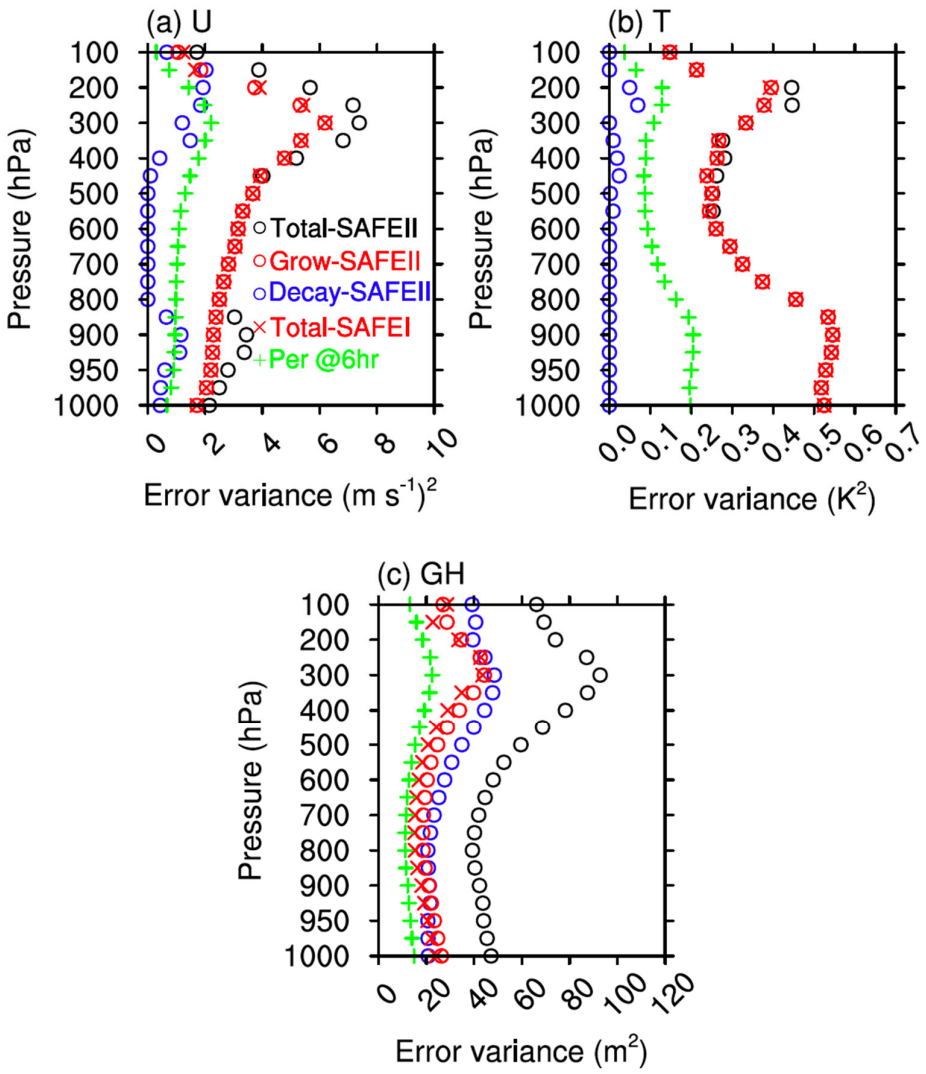
1074 Figure 8. Profile of the differences between the absolute fitting errors and the 95%

1075 confidence interval of perceived error variances at 1.5 days by SAFE-II (circle) and

1076 SAFE-I (cross) for variables (a) U, (b) T, and (c) GH for GFS-GSI operational forecast

1077 system.

1078



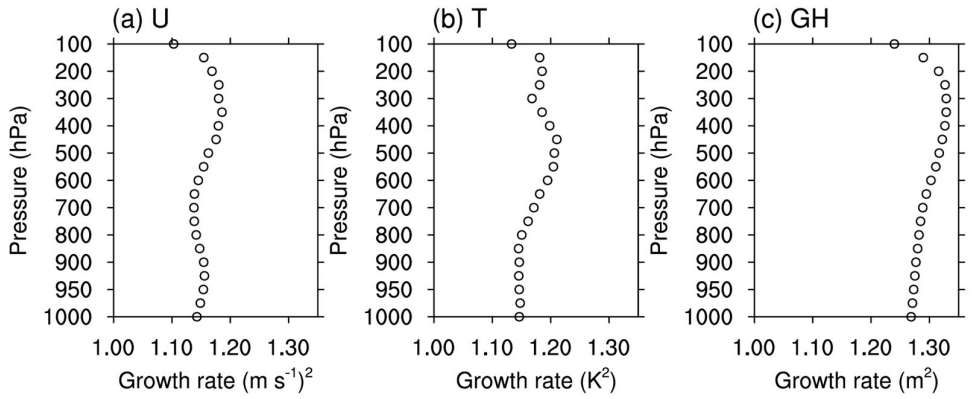
1079

1080 Figure 9. Same as Figure 8, but for estimated total (black circle), growing (red circle)

1081 and decaying (blue circle) analysis error variance by SAFE-II. Green and Red crosses

1082 represent the 6-hour perceived error variance and the estimated analysis error variance

1083 by SAFE-I, respectively.



1084

1085 Figure 10. Same as Figure 8 but for the estimated growth rate of error variance per 6

1086 hours.

1087

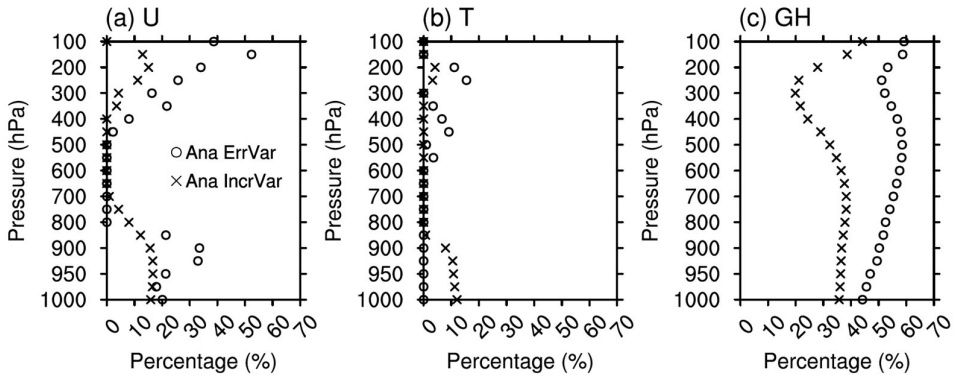


Figure 11. Same as Figure 8 but for the estimated percentage of decaying components in total analysis error variance by SAFE-II (circle) and variance of analysis increment (cross).