



Spatially extended estimates of analysis and short-range forecast error variances

By JIE FENG^{1*}, ZOLTAN TOTH¹ and MALAQUIAS PEÑA², ¹*Global Systems Division, ESRL/OAR/NOAA, Boulder, CO, USA*; ²*IMSG at Environmental Modeling Center, NCEP/NWS/NOAA, College Park, MD, USA*

(Manuscript received 22 December 2016; in final form 27 April 2017)

ABSTRACT

Accurate estimates of ‘true’ error variance between Numerical Weather Prediction (NWP) analyses and forecasts and the ‘reality’ interpolated to a NWP model grid (Analysis and true Forecast Error Variance, hereafter AFEV) are critical for successful data assimilation and ensemble forecasting applications. Peña and Toth (2014, PT14) introduced a Statistical Analysis and Forecast Error estimation (hereafter called SAFE) algorithm for the unbiased estimation of AFEV. The method uses variances between NWP forecasts and analyses (i.e. ‘perceived’ forecast errors) and assumptions about the time evolution of true error variances. PT14 successfully tested SAFE for the estimation of area mean error variances. In the present study, SAFE is extended by mitigating the effects of increased sampling noise and by accounting for the spatiotemporal evolution of forecast error variances, both critical for gridpoint-based applications. The enhanced method is evaluated in a Simulated Nature, Observations, Data Assimilation, and Prediction Environment using a quasi-geostrophic model and an ensemble Kalman Filter (EnKF). SAFE estimates of true analysis error variance are within 6% of the actual values, as compared to 24–55% deviations in EnKF estimates. The spatial correlation between estimated and actual true error variances was also found high (above 0.9) and comparable with EnKF estimates, but much higher than NMC method estimates (0.63–0.78). Estimates of the other two SAFE parameters, the growth rate and decorrelation of analysis and forecast error variances are within 3% of the corresponding actual values.

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

1. Introduction

Due to the chaotic nature of the atmosphere and to the presence of initial state and model related errors, despite continual improvements in NWP techniques numerical forecasts will never be perfect (Lorenz, 1963; Kalnay, 2002; Li and Ding, 2011). The accurate estimation of the variance between reality (interpolated onto the analysis/model grid) and the NWP analysis (i.e. true analysis error variance) or forecast states (true forecast error variance) is therefore critical for assessing the performance of analysis and forecast systems. In addition, the estimates of AFEV provide references for tuning initial ensemble perturbations (Toth and Kalnay, 1993, 1997; Molteni et al., 1996; Wei et al., 2008; Feng et al., 2014) and background error variances in data assimilation (DA) schemes (Fisher, 1996; Miyoshi and Kadowaki, 2008), respectively.

As reviewed by Peña and Toth (2014, hereafter PT14), the quality of NWP analyses is often assessed by either of two approaches. The first evaluates the analysis against observations valid at the same time (Houtekamer et al., 2005; Whitaker et al., 2008). This approach is limited by (a) the small number of

observations, (b) many of them used in making the analysis and therefore not being independent and (c) instrument and representativeness errors (with respect to the analysis grid) present in the observations (Dee and Silva, 1999). The second approach assesses analysis uncertainty through the DA schemes used in creating the analysis fields. These type of errors estimate rely on the same assumptions as the DA schemes and therefore cannot be considered independent estimates. DA-based approaches are computationally very expensive and due to approximations they may lead to biased or inaccurate analysis error variance estimates.

Most often the quality of numerical forecasts is assessed using either verifying observations or analyses as a reference for truth (Houtekamer et al., 2005; Whitaker et al., 2008). Such comparisons will be affected by errors present in either data source, especially at shorter lead times where the observational or analysis errors may have a magnitude comparable to that of the forecast errors. True forecast error variances can also be derived from methods primarily used to estimate background forecast error covariances in DA applications. The NMC method (Parrish and Derber, 1992, named after the National Meteorological Center

*Corresponding author. e-mail: jie.feng@noaa.gov

where it was developed, now called the National Centers for Environmental Prediction [NCEP]) uses differences between forecasts with different lead times all valid at the same time, as a surrogate for short-range forecast errors. Another approach uses the spread of ensemble forecast members (e.g. Evensen, 1994, 2003; Houtekamer et al., 1995; Fisher, 2003; Wei et al., 2010). Forecast error variance approximations with these methods critically depend on the choice of the lag between forecasts (NMC method) or of the initial ensemble perturbation variance, and therefore cannot provide impartial estimates.

PT14 introduced a Statistical Analysis and Forecast Error (SAFE) variance estimation method henceforth referred to as SAFE. SAFE uses perceived error (forecast minus verifying analysis) measurements and two assumptions to estimate AFEV. The first assumption is about the time evolution of the true forecast error variance, while the second is related to the correlation between errors in analyses and forecasts valid at the same time as a function of lead time. Note that the two assumptions are independent of those used in DA schemes and therefore SAFE can be suitably applied to evaluate and compare the performance of different DA-forecast systems. In a simple model environment, PT14 showed that SAFE analysis and forecast error variance estimates are statistically indistinguishable from the actual values given the sampling noise in the perceived error measurements. In addition, PT14 applied SAFE for the estimation of hemispheric mean error variances in forecasts from various NWP centres.

The present paper attempts to expand the study of PT14 in the following aspects: (1) To reduce the effect of sampling noise and improve accuracy, SAFE is upgraded with (a) a more suitable minimisation algorithm – the limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) algorithm (Byrd et al., 1995) in place of the Nelder–Mead Simplex method, and (b) additional measurements in the form of differences between forecasts valid at the same time (i.e. time-lagged forecast differences). (2) SAFE is applied and evaluated in a Simulated, Nature, Observing, Data Assimilation, and Prediction (SNODAP) environment with a numerical model of intermediate complexity instead of the Lorenz toy model. Such an environment where the AFEV are exactly known offers an ideal setting for a rigorous assessment of SAFE estimates. (3) Unlike the hemispheric mean estimates in PT14, SAFE will be applied for spatially extended error variance estimation, meeting real-world DA and ensemble forecast application needs. (4) To assess the relative merits of different approaches, SAFE estimates will be compared with error variance estimates from standard techniques such as ensemble Kalman filter (EnKF; e.g. Evensen, 1994, 2003; Whitaker et al., 2008) and the NMC method (Parrish and Derber, 1992; Wang et al., 2014).

The paper is organised as follows. The error estimation methodology and its enhancements are presented in Section 2. Section 3 describes the experimental environment for the implementation and testing of the method. Experimental results for

area mean and gridpoint error variance estimates are discussed in Sections 4 and 5, respectively, while a summary and conclusions are offered in Section 6.

2. Methodology

SAFE (PT14) uses a statistical model to describe the time mean evolution of measured variables (i.e. variances of perceived errors and lagged forecast differences) as a function of unknown variables (i.e. AFEV). The model is built upon prior knowledge regarding the evolution of (a) true forecast error variance, and (b) the correlation between true errors in a free forecast vs. in successive verifying analyses in a DA-forecast cycle, as a function of forecast lead time. The unknown parameters are estimated via minimising the fit of the statistical model to the measurements.

2.1. Decomposition of perceived error variances

As in any statistical minimisation problem, we seek to establish connections between measured and unknown variables. We define forecast lead time as $t_i = i \times \Delta t$, where i is the number, and Δt (12 h in this paper) the length of DA cycles. Let $\mathbf{F}_i$ and $\mathbf{F}_{i+l}$ be two forecasts with lead time indices i and $i + l$ ($l \times \Delta t$ lagged), respectively, both valid at the same time. The variance of the difference between lagged forecasts $\mathbf{F}_i$ and $\mathbf{F}_{i+l}$ can be written as:

$$d_{i,i+l}^2 = (\mathbf{F}_i - \mathbf{F}_{i+l})^2 = ((\mathbf{F}_i - \mathbf{T}) - (\mathbf{F}_{i+l} - \mathbf{T}))^2 = (\mathbf{x}_i - \mathbf{x}_{i+l})^2, \quad (1)$$

where $\mathbf{T}$ is the true state of the system valid at the same verifying time as the forecasts; and $\mathbf{x}_i = \mathbf{F}_i - \mathbf{T}$ and $\mathbf{x}_{i+l} = \mathbf{F}_{i+l} - \mathbf{T}$ are the true forecast errors at lead time indices i and $i + l$, respectively. The parentheses in Equation (1) denote spatial and temporal averages. Following the law of sum of variances, the right-hand side of (1) can be expanded to:

$$d_{i,i+l}^2 = x_i^2 + x_{i+l}^2 - 2\rho_{i,i+l}x_i x_{i+l}, \quad (2)$$

where x_i and x_{i+l} are the magnitudes of $\mathbf{x}_i$ and $\mathbf{x}_{i+l}$ in a L_2 norm; and $\rho_{i,i+l}$ is the correlation between $\mathbf{x}_i$ and $\mathbf{x}_{i+l}$. In Equation (2), x_i , x_{i+l} and $\rho_{i,i+l}$ are all unknown quantities to be statistically estimated. With the choice of $i = 0$, PT14 estimates the error variance x_0 in the analysis state $\mathbf{F}_0$ by comparing the measured perceived error variances with its statistically modelled values:

$$\widehat{d_{0,l}^2} = x_0'^2 + x_l'^2 - 2\rho_{0,l}'x_0'x_l', \quad (3)$$

where the hat represents the modelled value and the prime means the unknown parameters to be estimated. Varying l in Equation (3) yields a series of relationships between the measured and unknown variables. However, with each additional equation, new unknown variables (x_l' and ρ_l' – simplified from $\rho_{0,l}'$) are introduced beyond x_0' . To eliminate some of the unknown parameters, additional relationships need to be introduced.

2.2. Relationship between unknown parameters

To reduce the number of unknown parameters, PT14 assumed the time evolution of true forecast error variance follows well-established error growth patterns. As shown by Lorenz (1963), in the perfect model scenario, the initial, linear phase of error growth in chaotic systems follows an exponential curve:

$$x_l'^2 = x_0'^2 e^{-\alpha' \cdot l}. \quad (4)$$

with two independent parameters of the initial error variance $x_0'^2$ and the exponential growth rate α' . To describe error growth in the presence of nonlinear saturation, Lorenz (1982) introduced the logistic error growth model:

$$x_l'^2 = (S_\infty' \cdot c) / (e^{-\alpha' \cdot l} + c), \quad (5)$$

where $c = x_0'^2 / (S_\infty' - x_0'^2)$, and the saturation level S_∞' is the only additional parameter introduced beyond the two exponential growth parameters of x_0' and α' . Over areas with significant model drift-related errors (e.g. the tropics as convection is generally poorly parameterised), a general error growth model can be introduced to describe the evolution of the total error variance, including a model drift-related component (see Section 2.2.3 in Peña and Toth, 2014).

Considering the relationship between $\rho_1, \rho_2, \dots, \rho_l$ in versions of Equation (3), PT14 notes that ρ_1 measures how much the true first guess error $\mathbf{x}_1$ rotates to become the analysis error $\mathbf{x}_0$ as a consequence of observational information being ingested in one application of the DA scheme. Furthermore, PT14 postulates that with each subsequent application of the DA scheme, the analysis error vector is successively rotated by ρ_1 with respect to a free forecast in which the error evolves unperturbed according to the dynamics of the system, yielding the following relationship between true analysis and true forecast errors, all valid at the same time:

$$\rho_l' = \rho_1^l, \quad 0 < \rho_1 < 1. \quad (6)$$

Note that with increasing lead time, ρ_l' approaches zero as the true analysis and forecast error vectors become independent.

2.3. Cost function

2.3.1. Perceived errors. With the error growth and error decorrelation assumptions introduced above, the modelled temporal mean perceived error variance $\widehat{d_{0,l}^2}$ at lead time l can be expressed as:

$$\widehat{d_{0,l}^2} = x_0'^2 + x_l'^2 - 2\rho_1^l x_l' x_0', \quad (7)$$

where $x_l'^2$ is a function of the parameters of the selected error evolution model (Equation (4) or (5)). PT14 estimates the unknown parameters (x_0' , α and ρ_1) by solving the minimisation problem of the following cost function:

$$J = \max \left(\left| d_{0,l}^2 - \widehat{d_{0,l}^2} \right| \cdot w_{0,l}^{-1} \right). \quad (8)$$

To measure the difference between the observed ($d_{0,l}^2$) and modelled ($\widehat{d_{0,l}^2}$) quantities ($|\cdot|$ denotes the absolute value), PT14 opts to use the L_∞ norm that ensure a good fit over the entire range of l 's, rather than the more conventional L_2 norm. This approach minimises the maximal value (function $\max(\cdot)$) of the fitting errors at any lead time within the fitting period. The weights w are proportional to the uncertainty in the sample-based estimates of d (see also Section 2.3.3). For stable statistical estimates, the number of fitting terms in Equation (8) is chosen to be well above the number of unknowns. Since the initial transitional behaviour of analysis errors (Trevisan and Legnani, 1995) may deviate from exponential growth, only perceived errors with l greater than or equal to 2 will be used in this paper (i.e. perceived errors in 24-h or longer forecasts).

2.3.2. Lagged forecast differences. SAFE estimates of the unknown parameters are negatively influenced by sampling error in the time mean of individual perceived error measurements. To reduce the influence of sampling error, in addition to perceived errors used in PT14, lagged forecast differences will also be measured and modelled. In particular, it is assumed that in a perfect model environment used in this study, the variance between two lagged forecasts amplifies with the same exponential growth rate as the true forecast error variance (Lorenz, 1982). For example, the growth of variance between forecasts lagged 24 h apart can be modelled like:

$$\widehat{d_{i,i+2}^2} = d_{2,4}^2 \cdot e^{\alpha' \cdot l_{i-2}}, \quad i \geq 2. \quad (9)$$

Since no new unknowns appear in Equation (9), the corresponding fitting terms can be directly introduced into the cost function:

$$J = \max \left(\left| d_{0,l}^2 - \widehat{d_{0,l}^2} \right| \cdot w_{0,l}^{-1} \right) + \max \left(\left| d_{i,i+2}^2 - \widehat{d_{i,i+2}^2} \right| \cdot w_{i,i+2-1} \right), \quad (10)$$

$$l \geq 2, \quad i \geq 2.$$

2.3.3. Weights. Uncertainty in the sample-based estimates of time mean perceived error, or lagged forecast difference variances grows with increasing variance in these quantities as a function of increasing lead time. Correspondingly, the weights on fitting errors between the observed and estimated quantities at longer lead times must be reduced. PT14 defined the weights $w_{i,i+l}$ in Equation (10) as:

$$w_{i,i+l} = \text{SEM}_{i,i+l} / \sum_{i,i+l} \text{SEM}_{i,i+l}. \quad (11)$$

The sampling standard error of the mean (SEM) of perceived errors and forecast differences between lead time i and $i + l$ is given as follows:

$$\text{SEM}_{i,i+l} = sd_{i,i+l} \cdot f / \sqrt{N}, \quad (12)$$

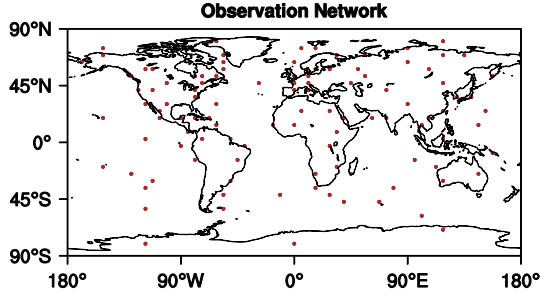


Fig. 1. Distribution of simulated observations (hollow circle).

where $sd_{i,i+1}$ is the sample standard deviation of $d_{i,i+1}^2$ in Equation (1), N is the sample size. $f = \sqrt{(1 + r_1)(1 - r_1)^{-1}}$ is an adjustment factor accounting for serial correlation in the sample (Bence, 1995), and r_1 is the lag-1 autocorrelation in the sample. Perceived error and lagged forecast difference terms in Equation (10) are given the same (unity) weight in the cost function.

2.3.4. Confidence interval. The SEM values can also be used to quantify confidence intervals for SAFE estimates. Assuming that the finite-sample mean of individual perceived error variance measurements follows a Gaussian distribution, the 95% confidence interval is defined by adding and subtracting 1.96 times the SEM value to/from the best estimates. In realistic situations, only SEM for perceived errors and forecast differences can be computed. In the present idealistic SNODAP study, however, the individual true forecast error variance and analysis – forecast correlation values are also known. Therefore, confidence intervals for their time mean estimates can and will also be computed.

2.3.5. Minimisation. The minimisation of the cost function defined by Equation (10) is a nonlinear constrained optimisation problem. Since the Nelder–Mead simplex minimisation method (Lagarias et al., 1998) used in PT14 is sensitive to the choice of the first guess parameters, in this paper the L-BFGS algorithm (Byrd et al., 1995) will be used. This algorithm is designed for the solution of constrained nonlinear optimisation problems and can be easily ported across different computational platforms.

3. Experimental design

In this paper, the SAFE method, as described and enhanced in Section 2 will be tested and evaluated in an intermediate complexity perfect model SNODAP environment. In a SNODAP, a numerical model integration (i.e. a ‘free’ forecast) is considered as reality. To focus on chaotically growing, and avoid model-related errors, the global quasi-geostrophic model of Marshall and Molteni (1993 – MM-QG) with a three-layer vertical pressure coordinate system (200, 500, 800 hPa) and at T21 spectral

resolution (truncated at wavenumber 21) will be used both for simulating and predicting reality. The forcing in MM-QG is carefully chosen in the form of spatially varying but temporally stationary potential vorticity source terms so that the observed large-scale extratropical boreal winter climatology is reasonably reproduced. MM-QG is well suited for this study as with its 6144 total degrees of freedom it is complex enough to represent basic baroclinic processes critical to extratropical forecast error growth. The lack of convective processes, however, renders the model’s tropical circulation less realistic.

After discarding the first 500 days of integration started from an operational analysis field, geopotential height (GH) at 12-h intervals over a 90-day period is used to simulate reality. Sites for simulated observations are selected to mimic the geographical distribution of real-world traditional *in situ* observing networks (Fig. 1). Observations are simulated by adding Gaussian uncorrelated noise with a variance of 260, 80 and 30 m^2 at the 3 model levels (from top to bottom) to simulated reality. To assimilate the simulated observations, the EnKF scheme described in Appendix A is used. Vertically correlated noise (see Appendix A) is superposed on the true state to produce an initial first guess. Analysis, 12-h first guess as well as 30-day forecasts are created twice a day with the same MM-QG model used to simulate reality. This, perfect model scenario ensures that all forecast error originates from initial condition (i.e. analysis), and not model-related errors. After the behaviour of the DA – forecast system asymptotes (30 days) following the introduction of an arbitrary initial first guess, a total of 180 consecutive cases are selected for experimental use in this study.

As reality is perfectly known, all components of the SNODAP can be perfectly diagnosed. This offers an opportunity to assess the quality of AFEV estimates produced by SAFE in a realistic setting using perceived error measurements by comparing such with actual error variances. Beyond such estimates simulating real-world applications, some components of SAFE will also be tested in an ‘ideal’ setting. Error growth curves, for example, can be fitted to actual error variances to ascertain the validity of some assumptions used in SAFE (see Section 5.2) – something that can only be accomplished in SNODAP experiments.

4. Spatial mean error variance estimates

In this section, the method described in Section 2 will be used and evaluated for the estimation of area mean forecast error variances. The estimates will be based on perceived error measurements with the assumption of either exponential (limited to short-range error growth) or logistic error growth assumptions.

Compared to the method used in PT14, the version of the SAFE method proposed in Section 2 in this study is enhanced by the use of an additional, lagged forecast difference term in the cost function. The effect of this change on the accuracy of SAFE estimates based on short-range perceived error measurements

Table 1. Estimated values of parameters x_0^2 , α and ρ_1 with the Statistical Analysis and Forecast Error (SAFE) estimation in PT14 and its enhanced version proposed here for the 500 hPa global (GB), Northern (NH) and Southern (SH) Hemisphere, and Tropical belt (TRO) regions. The actual values of the parameters and the SEM corresponding to the 95% significance level are also shown.

		GB	NH	SH	TRO
x_0^2	Actual/1.96SEM	42.23/2.43	32.12/2.78	72.82/5.87	25.32/1.30
	New estimate	39.58	34.04	68.16	19.51
	PT14 estimate	44.10	39.27	77.72	15.23
α	Actual/1.96SEM	0.405/0.024	0.574/0.033	0.297/0.024	0.300/0.027
	New estimate	0.408	0.565	0.287	0.331
	PT14 estimate	0.391	0.540	0.270	0.389
ρ_1	Actual/1.96SEM	0.84/0.010	0.789/0.021	0.859/0.014	0.860/0.01
	New estimate	0.837	0.789	0.838	0.844
	PT14 estimate	0.857	0.822	0.871	0.790

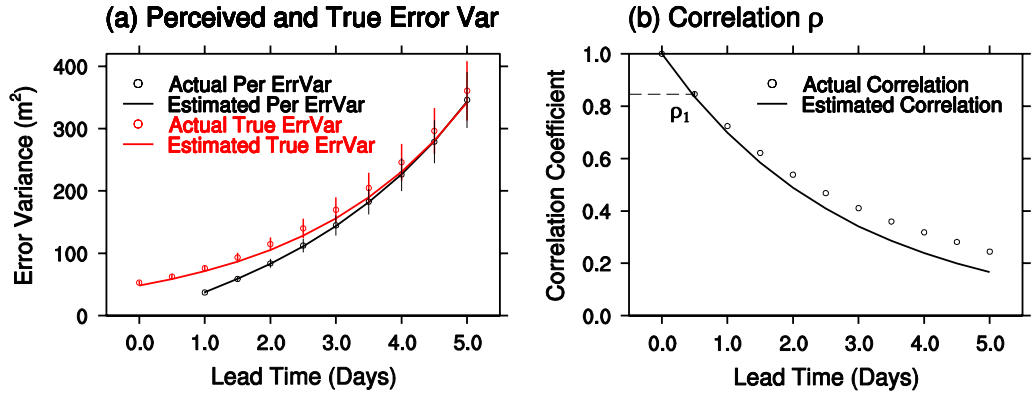


Fig. 2. (a) Globally (vertically and horizontally) averaged 90-day time mean perceived (black) and true (red) forecast error variances, and (b) resulting correlation of true forecast and analysis errors as a function of lead time. Hollow circles and lines represent actual and estimated values (assuming exponential error growth), respectively. Vertical bars show the 95% confidence interval corresponding to sampling uncertainty in the time mean error estimates. Since ρ_i ($i > 1$) is a simple function of ρ_1 , a confidence interval is shown only for ρ_1 . The estimated and actual values of the unknown parameters are $x_0^2 = 48.12$, $\alpha = 0.392$, $\rho_1 = 0.836$ and 53.0 , 0.38 , 0.85 , respectively.

with an exponential error growth assumption is shown in Table 1. In 11 out of the 12 regional parameters presented in Table 1, the new method provides better estimates while in the majority of the cases the new, but not the PT14 estimates are within the 95% sampling confidence interval of the actual values.

4.1. Exponential error growth

First, we test the applicability of the exponential error growth assumption. Assuming that the effect of nonlinear saturation on short-range forecast errors is small, we use perceived error measurements in the 1–5 day lead time range (24-, 36-, ..., 120-h). Globally (both vertically and horizontally) averaged simulated and actual perceived and true error variances, as well as the correlation between analysis and forecast errors valid at the same time, along with their 95% confidence intervals are shown in Fig. 2. Importantly, all simulated perceived error values fall within the 95% confidence intervals of the corresponding 1–5 day lead time measurement-based perceived error values used in the estimation. This indicates that the assumptions used in SAFE are consistent with the measurements from the SNOD-AP analysis-forecast system.

The estimated true error variance values also fall within the 95% confidence interval of the actual true error variance values in the 1–5 days lead time fitting period. Larger deviations are observed at initial and 12-h lead time between the extrapolated estimated and the actual error variances, presumably due to the assumption that all initial errors grow exponentially. Note the significant, up to a factor of two differences between the perceived and true forecast error in the first 2–3 days lead time range, attesting to the importance of properly considering analysis error variance when evaluating short-range forecast performance.

If we compare the equations for actual (x_t^2) and perceived ($x_0^2 + x_t^2 - 2\rho_{0,t}x_t x_0$) forecast error variances, we find that their difference is mainly determined by the parameter $\rho_{0,t}$ which is related to the size of change the DA makes to the first guess forecast. Specifically, if the observing network contains large amount of available observational information, the analysis error would be extensively rotated from the first guess error by the observations used in DA cycles, generally resulting in a small ρ_1 . The perceived error variance is likely to be larger than the true forecast error variance within the first few days. In contrast, if the observing network has only limited information,

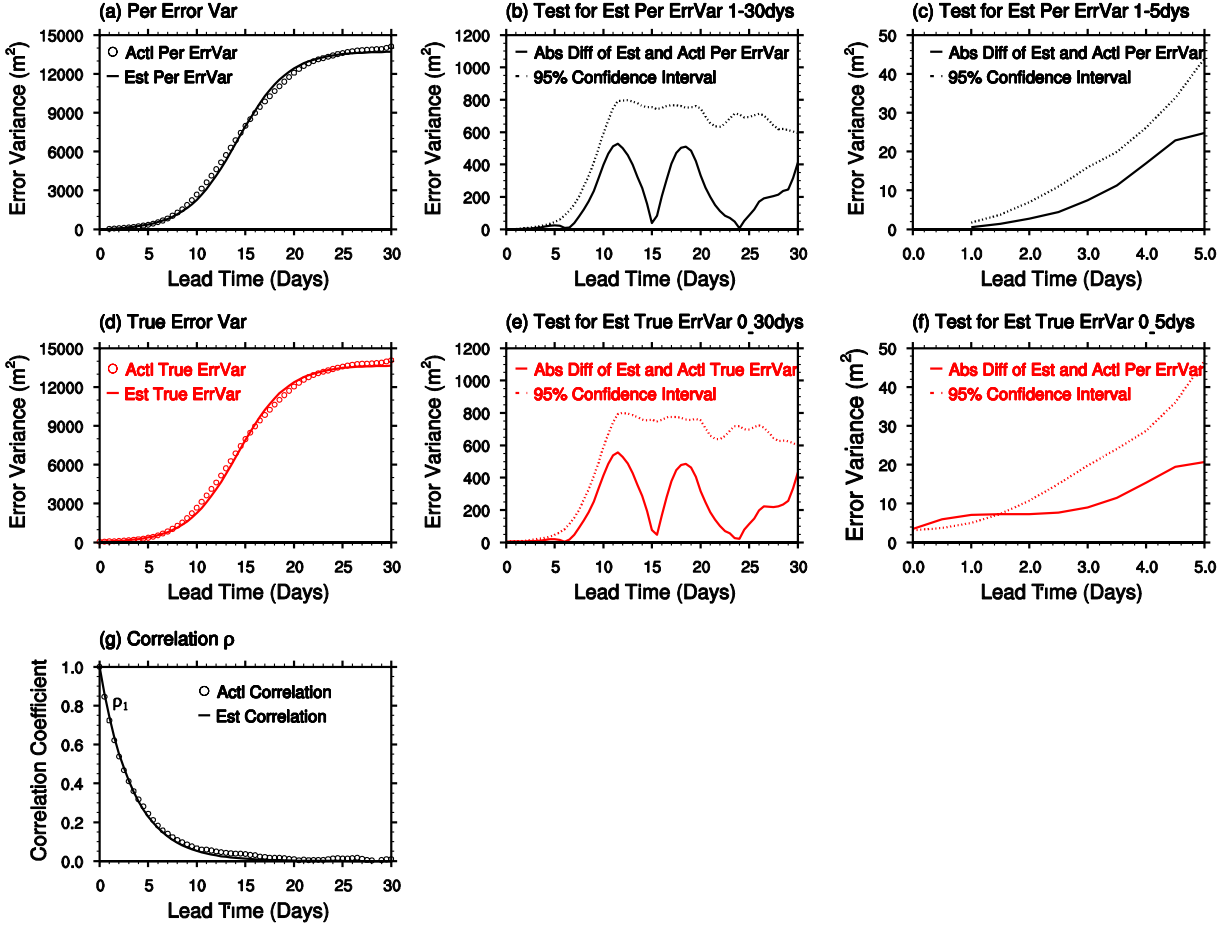


Fig. 3. Globally (both vertically and horizontally) averaged 90-day time mean actual (open circles) and simulated (assuming logistic error growth, continuous lines) perceived (a) and true (d) forecast error variances, and their absolute differences (continuous lines for perceived (b and c, the latter a zoom-in version of b), and true error (e and zoom-in f), as well as the 95% confidence level corresponding to sampling uncertainty in the time mean estimates (dotted curves). The correlation between true analysis and forecast errors is shown in (g). The estimated and actual values for the known parameters are $x_0^2 = 56.5$, $\alpha = 0.38$, $\rho_1 = 0.86$, $S_\infty = 13672.8$, and $53.0, 0.38, 0.85, 13889.1$, respectively.

ρ_1 could be relatively large, and hence the true forecast error variance would be more likely to exceed the perceived error variance. For longer lead times, due to the decrease in $\rho_{0,l}$ with increasing lead time, the analysis error becomes decorrelated from the true forecast error, and the perceived error variance would be simply the sum of AFEV theoretically.

4.2. Logistic error growth

Beyond an initial, quasi-linear phase of exponential growth, due to the finite size of dynamical systems the evolution of forecast errors is also influenced by nonlinear saturation (Lorenz, 1965). To describe such behaviour, we use the logistic error growth model with perceived error measurements in the 1–30 day lead time range. Attested by Fig. 3 and its legend, three of the four estimated values have less than 2% error when compared to their actual counterpart, while the variance/absolute value of the estimated analysis error variance is within 7/4% of the

actual values. Attested by Fig. panels 3a and d, SAFE can reasonably simulate both the perceived and the true forecast error variance. In fact, the fitting error is lower than the 95% sampling uncertainty confidence level at every time level from 0 to 30 days, except where true errors are extrapolated to the 0–12 h lead time period (Fig. panels 3b, c, e, and f). Not surprisingly, the use of the more general, logistic error growth model and more measurements leads to more informative (i.e. with an estimate of the saturation value) and slightly more accurate (i.e. improved analysis error) estimates when compared to those with the exponential growth assumption (cf. Figs. 2 and 3).

5. Spatially extended error variance estimates

In this section, we extend the application of SAFE from spatial mean (PT14 and Section 4) to pointwise error variance estimation.

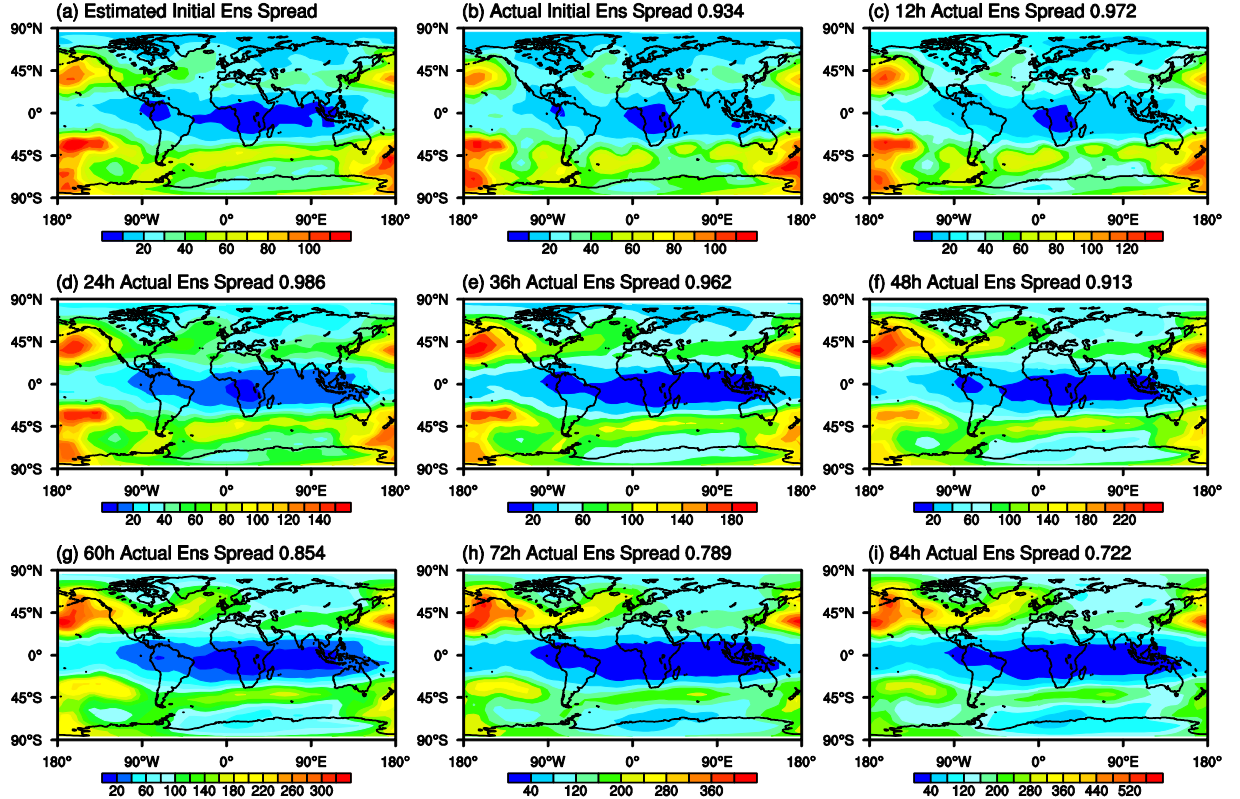


Fig. 4. Distribution of 90-day sample mean estimated (a) and actual (b) 500 hPa initial ensemble spread and actual 12–84-h ensemble spread. The spatial correlations of panels (b–i) to (a) are listed above the panels.

5.1. Methodological considerations

Gridpoint as opposed to large area mean estimates of AFEV are potentially more useful in practical applications like the specification of initial perturbation spread in ensemble forecasting or background error variance in DA systems. Gridpoint, as compared to area mean estimation, however, has some special challenges. First, the lack of spatial averaging of the input (perceived error) measurements results in increased sampling noise. And second, one must recognise that forecast errors verified at a selected location originate from an upstream location.

5.1.1. Sampling noise reduction. While analysis error variance (x_0^2) and error growth (α) estimates even averaged over large domains show large, greater than a factor of two variations (see Table 1), the correlation between analysis and first guess forecasts (ρ_1) exhibit much less (less than 10%) variability. To reduce the effect of sampling noise in the estimation of ρ_1 while retaining detailed information in the gridpoint estimates of the geographically more variant x_0^2 and α , in the gridpoint estimation of the unknown parameters (Section 5.3), we prescribe ρ_1 to its domain average estimates listed in Table 1. Note that the gridpoint estimates of x_0^2 and α are insensitive to the particular

choice of a spatial filter used (i.e. a broad Gaussian filter or domain means as in Table 1).

5.1.2. Spatiotemporal evolution of errors. Analysis error variance (x_0^2) is used in SAFE in two different contexts: as the error in the initial condition of NWP forecasts (Equations (4) and (5)), and as errors in the analysis used to verify the NWP forecasts (Equation (7)). When SAFE is applied over large latitudinal belts as in PT14 and Section 4, these two quantities can be considered identical. In gridpoint applications, the initial errors affecting forecasts of various lead times at a selected verification location originate, however, from different, upstream locations. SAFE analysis and growth rate estimates derived through the use of perceived forecast errors of different lead times at a selected verification gridpoint therefore are misplaced due to the spatiotemporal evolution of different lead time forecast errors and reflect conditions upstream of the verification gridpoint.

Bishop and Toth (1999) proposed the use of ensemble perturbations to estimate the spatiotemporal evolution of analysis – forecast errors. Given the multiple lead time and large number of gridpoints involved in SAFE, the application of their Ensemble Transform method, however would be algorithmical-

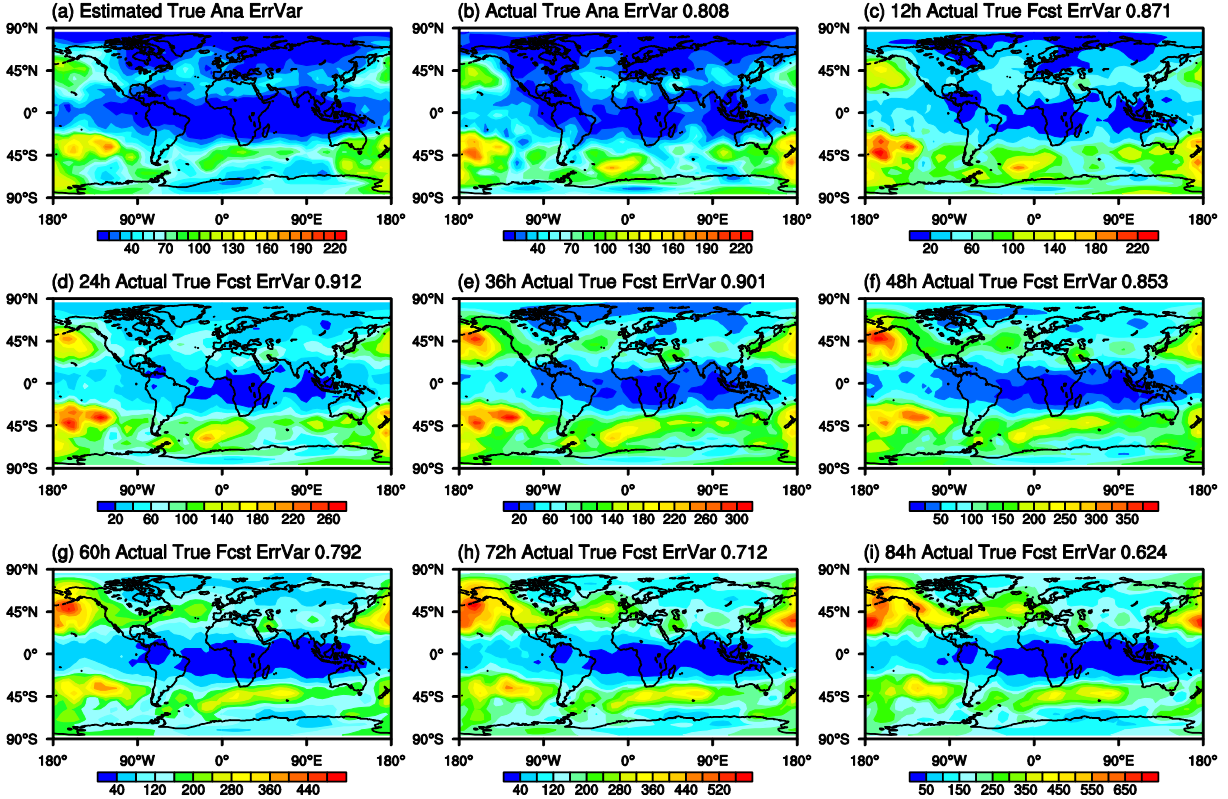


Fig. 5. Distribution of 90-day sample mean estimated (a) and actual (b) 500 hPa analysis error variances and 12–84-h actual true forecast error variances. The spatial correlations of panels (b)–(i) to (a) are listed above the panels.

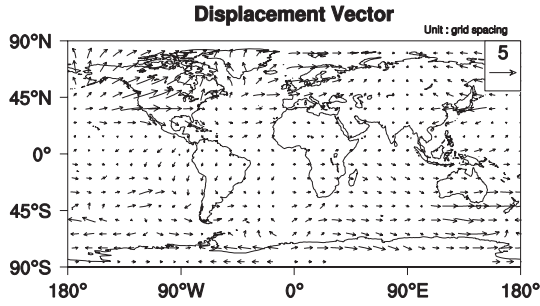


Fig. 6. Estimated displacement vector field transposing Fig. 4b to a.

ly complex and lead only to approximate results. Instead we substitute forecast error variance with ensemble perturbation variance¹ and use the SAFE method to simulate and quantify the misplacement of SAFE analysis error variance estimates. In particular, an exponential growth curve (Equation (4)) is fit to ensemble variance values measured at different lead times at each gridpoint. Using ensemble variances from the same 1–5 days lead-time range that is used in the error estimation, we ensure that the spatiotemporal behaviour of ensemble perturbation variances will well capture that of the true analysis – forecast errors manifested in SAFE.

The initial ensemble perturbation variance estimated via fitting 1–5 day ensemble forecast perturbation variance at each gridpoint is shown in Fig. 4a. While the estimated initial ensemble perturbation variance field is strongly correlated with the actual initial ensemble perturbation variance (Fig. 4b, 0.934), the estimated field is clearly misplaced. Larger than actual estimated values along the West Coast of North America (NA), for example, reflect the influence of larger initial ensemble variance over the Pacific propagating and growing into the coastal area where the spread influenced by those initial conditions are evaluated. Interestingly, the estimated initial ensemble perturbation variance has the highest correlation with 24-h ensemble forecast perturbation variance (0.986, see Fig. 4d), indicating that the misplacement of the SAFE gridpoint initial variance estimates may best correspond with an approximately one-day evolution of the errors.

To confirm that ensemble perturbation variance is a good predictor for the spatiotemporal evolution of true analysis – forecast error variance, analogous to Fig. 4, in Fig. 5 we show the evolution of the global distribution of the SAFE estimate of true analysis error variance, along with the actual AFEV. Apart from a significant disparity in their magnitudes, the overall pattern of error variance evolution is qualitatively well captured by

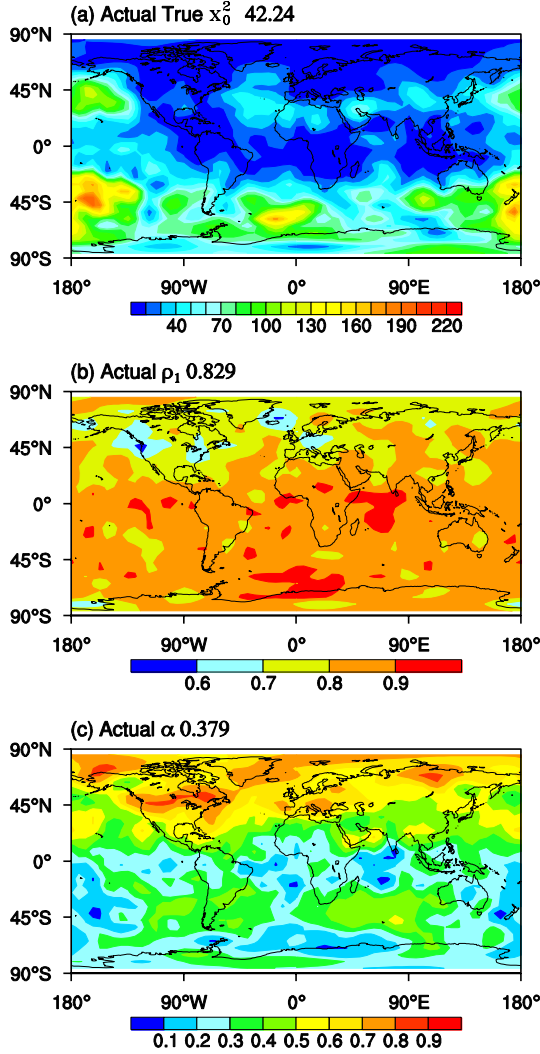


Fig. 7. Distribution of the actual values for the 500 hPa parameters (a) analysis error variance x_0^2 , (b) correlation of analysis and background forecast errors ρ_1 and (c) exponential growth rate. Global mean values are 42.24, 0.829 and 0.379, respectively.

the ensemble variance (cf. Fig. 4). Note that like in the ensemble, the estimated initial error variance reassuresly correlates best (though at a lower, 0.912 level due to more prevalent noise in a single error realisation as compared to the statistically more robust ensemble variance estimates) with the one-day forecast error variance.

Thus the evolution of ensemble (cf. Fig. 4a and b) can be used to describe the spatial mismatch between Fig. 5a and b. To quantify the displacement in SAFE analysis error variance estimates, we use the field alignment (FA) method of Ravela (2007) and Ravela et al. (2007). In our application, the displacement vector field (DV, see Fig. 6) is determined by a variational algorithm that minimises the difference between trial fields – *actual* initial ensemble variance estimates shown in Fig. 4b transposed by a DV field iteratively estimated by FA, and the

initial ensemble variance field *estimated* by SAFE (Fig. 4a). The large amplitude vectors over the NW part of NA reflect the previously noted fast propagation of initial errors from the NE Pacific to NA, while small vector amplitudes for example over Asia are indicative of mostly local error development. The final gridpoint estimates of analysis error variance evaluated in Section 5.3 will be transposed by the vectors opposite to those shown in Fig. 6.

5.2. Distribution of actual parameters

An advantage of working in a SNODAP environment is that all information about the natural and NWP systems is either readily available or easily attainable. Before applying SAFE to estimate its three free parameters, in this section we present their actual values. The analysis error variance (x_0^2 , Fig. 7a) is calculated at each gridpoint as the time mean of the variance between the EnKF analysis fields and the MM-QG model integration considered reality in SNODAP. The correlation between analysis and forecast errors (ρ_1 , Fig. 7b) is calculated at each gridpoint as the temporal correlation between the series of actual true analysis and 12-h forecast error variances valid at the same time.

The third unknown parameter of SAFE, the exponential growth parameter (α , see Fig. 7c), is determined at each gridpoint by fitting the exponential growth curve of Equation (4) to actual true forecast errors in the 1–5 days lead time range. Note that strictly speaking, due to the spatiotemporal propagation of errors discussed in Section 5.1.2, α computed here is representative of error growth *upstream* of the gridpoint where it is displayed. Such a spatial mismatch, however, will not affect the evaluation of growth parameter estimates in the next section as due to the use of the same 1–5 days lead time fitting range, the SAFE estimate of α is affected by exactly the same spatial misplacement. A comparison of the fitting errors (i.e. difference between actual and modelled true error values, not shown) and the corresponding SEM values (see Equation (11)) reveal that the exponential error growth model is fully consistent with the experimental data.

The larger growth rates observed over the Northern (NH) compared to the Southern Hemisphere (SH) extratropics in Fig. 7c are due to the choice of boreal winter forcing in the MM-QG model. Note the general tendency for analysis error variance (Fig. 7a) to be lower over well observed and low growth rate areas (e.g. N Europe, cf. Figs. 1 and 7c). The generally lower analysis error variance over the NH also indicates that the denser observing network over the NH more than compensates for the error amplifying effect of larger growth rates there. Due to the presumably larger changes introduced by more voluminous observational data over the NH, the correlation between the error in the concurrent background forecast and analysis fields is lower there (Fig. 7b). The ideal parameter values displayed in Fig. 7 will be used in the next section as a reference when evaluating realistic perceived error-based SAFE estimates.

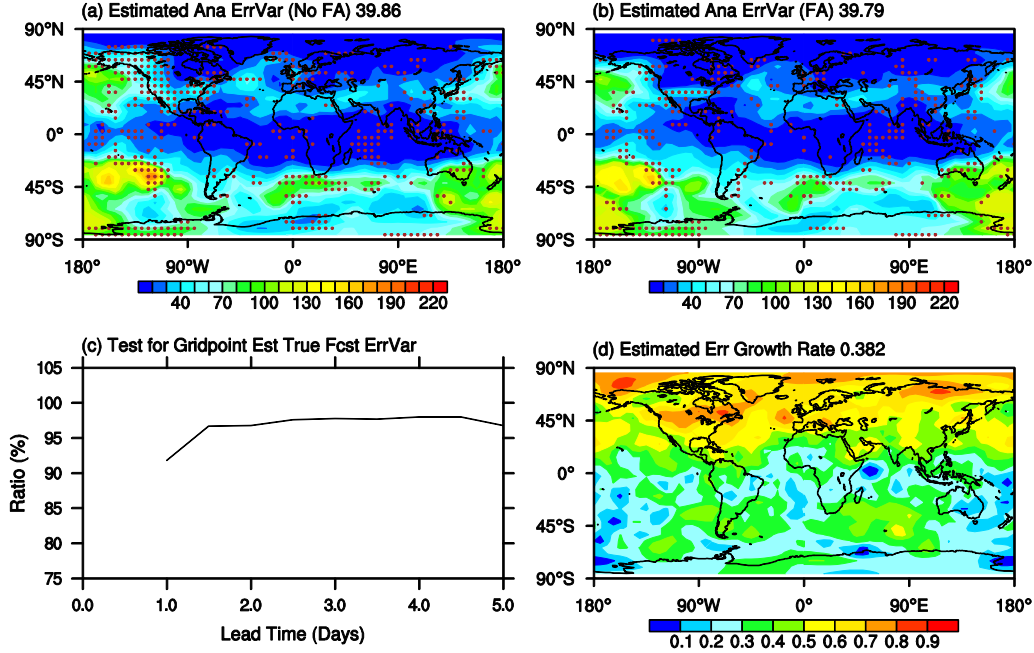


Fig. 8. Distribution of estimated 500 hPa analysis error variance without (a) and with (b) the application of field alignment (FA) technique, and the estimated exponential error growth rate (d). Global mean value of (a), (b) and (d) are 39.86, 39.79 and 0.382, respectively. (c) Percentage of gridpoints with estimated true forecast error variance within the 95% sampling confidence interval of their respective actual time mean values. For further details, see text.

5.3. Estimated parameters

As described in Section 5.1.1, in the gridpoint application of SAFE ρ_1 is prescribed from prior area mean estimates over the NH and SH extratropics and the Tropical belt (Table 1). Gridpoint perceived errors are statistically modelled and fitted to the time mean actual perceived errors in the 1–5 days lead time range. The fit of the modelled time mean perceived error variance at each gridpoint is within the 95% sampling uncertainty confidence interval around the actual time mean perceived error variance values at each time level.

Fig. 8a and d shows the distribution of the estimated true analysis error variance (x_0^2) and growth rate (α), respectively. The global mean of the estimated gridpoint x_0^2 and α values (39.86 and 0.382) are close to their actual values (42.24 and 0.379, cf. Figs. 7 and 8). Another measure of performance, the spatial correlation between the estimated and actual fields for x_0^2 and α is 0.81 and 0.85, respectively.

The percentage of gridpoints where the estimated true error value falls within the 95% sampling uncertainty confidence interval around the actual time mean true error value is shown in Fig. 8c. While the forecast error estimates over the fitting period of 1–5 day lead time range are within the confidence interval at 96–97% of the gridpoints (except at day 1 lead time where this condition is met at 92% of the gridpoints), the analysis error variance extrapolated from the fitting period with the use of the exponential error growth relationship falls out of the

confidence interval at 23% of the gridpoints (marked by brown dots on Fig. 8a). Conspicuously, points with poor estimates tend to be associated with areas of large DV magnitudes (cf. Figs. 8a and 6), indicating that the displacement of pointwise error estimates discussed in Section 5.1.2 has a clear negative effect on the quality of analysis error variance estimates.

When the position of true analysis error variance estimates (Fig. 8a) is corrected with the FA method using the inverse of the vectors derived from ensemble variances (Fig. 6), the percentage of gridpoint estimates that fall outside the sampling uncertainty confidence interval around the actual true error is reduced from 23 to 19% while the spatial correlation between the estimated and actual true analysis error variance fields increases from .81 to .90² (see Fig. 8b).

5.4. Comparison with other methods

In this section, AFEV estimates with SAFE are compared with those from other widely used methods. In Table 2, first we compare the performance of perceived error variance in predicting the magnitude and spatial distribution of true error variance. As seen from Table 2 perceived errors, due to the use of analyses as verifying fields, seriously underestimate true forecast error variance at early (shorter than 48 h) lead time ranges. Beyond 48 h, perceived error and SAFE estimates perform comparably. As for predicting the spatial distribution of the true forecast error variance, perceived errors and SAFE have comparable

Table 2. Spatial correlations between actual perceived and true forecast error variances at different lead times (2nd row). The spatial correlations between estimated (with SAFE algorithm) and actual true forecast error variances are shown on the 3rd row. Their global mean error variances at different lead times are listed on 4th–6th rows, respectively. The initial perceived errors are 0 and thus its field is replaced by the 12-h field when calculating the spatial correlation.

		0 h	12 h	24 h	36 h	48 h	60 h	72 h	84 h	96 h
Corr	SAFE	0.90	0.87	0.91	0.94	0.95	0.96	0.97	0.97	0.97
	PerErr	–	0.88	0.91	0.93	0.94	0.96	0.97	0.98	0.98
Error Var (m ²)	SAFE	40.0	47.5	56.8	68.3	82.6	100.6	123.4	152.3	189.3
	PerErr	–	15.0	30.9	49.4	71.1	96.3	125.0	158.8	198.2
	Actual	42.2	50.3	61.7	76.6	95.0	116.9	143.0	174.2	211.9

Table 3. Comparison of EnKF, NMC and SAFE estimates of the actual 500 hPa analysis and first guess forecast error variance fields using global mean magnitude and spatial correlation.

		EnKF	NMC	SAFE
Actual analysis error variance (m ²): 42.2	Spatial corr	0.92	NA	0.90
	Error variance(m ²)/deviation of Est	19.0/55%	NA	39.8/6%
Actual first guess error variance 12 h (m ²): 50.3	Spatial corr	0.91	48–24 h: 0.63 24–12 h: 0.78	0.87
	Error variance(m ²)/deviation of Est	Before inflation: 22.8/55% After inflation: 38.6/23%	48–24 h: 48.9/3% 24–12 h: 18.9/62%	47.5/6%

performance. Since perceived errors are not available at analysis (i.e. zero lead) time, we tested the use of 12-h lead time perceived error variance for the prediction of the distribution of analysis error variance. Its correlation with the actual true error, 0.81 (the highest correlation value for perceived error variance at any lead time, not shown in Table 2) is notably below the 0.90 correlation for SAFE estimates. These results suggest that as expected, SAFE may have the largest improvement upon perceived error estimates at the analysis and short lead times.

As one of a variety of ensemble-based DA methods, EnKF offers error variance estimates for both analyses and forecasts, while the NMC method for estimating background error covariance (Appendix B) addresses only background forecast error variances (see Table 3). As for the absolute magnitude of analysis and background forecast error variances, EnKF suffers from a serious underestimation of analysis (55%) and forecast (55% before and 23–24% after the application of covariance inflation) error variances. Depending on the choice of forecast lead times, the NMC method³ underestimates the actual forecast error variance by 3% (48- and 24-h) or 62% (24- and 12-h lagged differences), compared to a 6% error in the SAFE estimate.

Despite the large errors in its magnitude estimates, EnKF can well predict the spatial distribution of error variances (.92 and .91 correlation between the actual and predicted analysis and forecast error variance fields, respectively). Here, EnKF is used to predict error variances in EnKF analyses or background forecasts. It is not clear how much the performance may degrade had it been used to predict error variances in another DA scheme, like the hybrid Gridpoint Statistical Interpolation (GSI) operational at NCEP. The NMC method is less skillful

with a .63 (48- and 24-h) and .78 (24- and 12-h lagged differences) correlation between the actual and its predicted spatial background error variance distributions. These results can be compared to .90 and .87 correlations between the spatial distribution of actual and SAFE estimated AFEV, respectively.

6. Conclusions and discussion

Accurate estimation of true analysis and short-range forecast error variances are critical to the proper evaluation and calibration of DA and forecast systems, including the optimal tuning of background forecast error and initial ensemble perturbation variances. To advance error variance estimation, the SAFE method introduced in PT14 was further developed and tested in this paper. Barring systematic errors, the SAFE assumes that the true initial error variance (x_0^2) amplifies in the forecast according to well-established exponential or logistic error growth relationships. Assuming also that errors in the verifying analysis decorrelate exponentially with increasing lead time forecast errors, perceived error variance (variance between forecasts and verifying analyses) is decomposed and related to the unknown variables of true analysis error variance (x_0^2), dynamical growth rate (α), and the correlation between analysis and background forecast errors (ρ_1). The unknown variables are determined by minimising the difference between the modelled and sample-based measurements of perceived error variance at different lead times.

To improve the performance of SAFE, we introduced additional constraints into the cost function (i.e. the variance between lagged forecasts at different lead times), as well as

switched to the use of the L-BFGS minimisation algorithm that is more robust to the choice of an initial guess. For gridpoint applications more affected by sampling noise, we introduced area mean estimates of the slowly varying ρ_1 parameter as a constraint in SAFE. Also, the geographical placement of pointwise error variance estimates was corrected using ensemble forecast perturbations as a proxy for the spatiotemporal evolution of error variance.

The modified SAFE method was evaluated using Marshall and Molteni's (1993) QG model in a perfect model Simulated Nature, Observations, Data Assimilation, and Prediction (SNODAP) Environment where reality is perfectly known. Simulated area mean perceived error variance was found to be within sampling uncertainty of sample-based measurements, indicating that the SAFE statistical model is consistent with the observations. 500 hPa height SAFE area mean analysis error variance, exponential growth rate and error correlation estimates were also close (i.e. 6% or less deviation) to their actual values. This favourably compares with error variance estimates with up to 60% deviations by alternative methods, such as the EnKF and the 'NMC' methods. With a logistic error growth equation, the nonlinear saturation level of forecast error variance was also accurately (with less than 2% deviation) estimated.

Perceived error variance provides as good of an estimate of the *distribution* of the true error variance field as the SAFE estimates. However, SAFE provides a better prediction of error variance distribution at analysis time, and much better estimates of the absolute value of error variance for lead times shorter than 48 h. SAFE and EnKF both showed high skill in predicting the spatial distribution of analysis error variance (with spatial correlations above 0.9 between predicted and actual values), though it is not clear how skillful EnKF may be in predicting the error distribution in analyses other than those produced by EnKF itself that were evaluated in this study, such as from the GSI from the NCEP.

When the model-related errors are negligible, the growth of true forecast error variance could be simulated by the exponential or logistic equation for short- or long-range forecasts, respectively. Peña and Toth (2014) also introduced a general error growth model to describe the error growth when significant model drift is present. Such an approach can facilitate the application of SAFE in an operational NWP environment. On the other hand, the SAFE method assumes that all initial errors project onto the subspace of exponentially fast growing perturbation directions. Analysis errors, however, have non-growing components as well (Errico and Langland, 1999; Hamill and Whitaker, 2011; Privé et al., 2013). The neglect of decaying errors may lead to an underestimation of analysis error variance, with relatively poor performance (5% deviation from actual value, Table 1) than the estimates of other parameters, including forecast error variances at and beyond 24 h lead time (less than 3% deviation). Beyond potential enhancements to de-

scribe error behaviour in the short, transitional period affected by decaying errors, future studies will explore the use of the enhanced SAFE method in estimating errors in forecasts made with imperfect models at various NWP operational centres.

Acknowledgements

Scott Gregory and Dr. Isidora Jankov provided the field alignment software package (Peña et al. 2016) used in this study. Discussions with Drs. Hongli Wang (Panasonic Corporation, formerly GSD) and Yuanfu Xie (GSD), Krishna Kumar (NCEP) and Jordan Alpert (NCEP) are also acknowledged. The support and encouragement of Kevin Kelleher, Director of GSD, and Dr. Lidia Cucurull, Chief of the Global Observing Systems Analysis Group, GSD, are gratefully acknowledged. Dr. Brian Etherton of GSD provided valuable comments on an earlier version of this manuscript. This research was performed while the first author held a National Research Council Research Associateship award at GSD, ESRL/OAR/NOAA.

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes

1. Even though both in our simulated or in realistic environments initial ensemble variance only approximates true analysis error variance, as we will see below, the dynamics of error propagation can be well approximated from the ensemble.
Note that even in a perfect model environment, ensemble perturbation variance would provide a perfect prediction of the forecast error variance only if the ensemble initial variance perfectly matches analysis error variance, which in practice is unattainable.
2. The minor change in the global mean of analysis error variance estimates with FA (39.79) compared with the original estimate of 39.86 (cf. Fig. 8d and a) is due to a modest deformation of the estimated analysis error variance field associated with the application of FA.
3. Note that the NMC method provides estimates for background forecast but not analysis error variances.
4. Smaller inflation numbers result in filter divergence while larger numbers lead to suboptimal performance.

References

- Bence, J. R. 1995. Analysis of short time series: correcting for autocorrelation. *Ecology* **76**(2), 628–639.
- Bishop, C. H. and Toth, Z. 1999. Ensemble transformation and adaptive observations. *J. Atmos. Sci.* **56**, 1748–1765.
- 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.

- Dee, D. P. and Silva, A. M. D. **1999**. Maximum-likelihood estimation of forecast and observation error covariance parameters. Part I: methodology. *Mon. Wea. Rev.* **127**(8), 1822–1834.
- Errico, R. M. and Langland, R. **1999**. Notes on the appropriateness of ‘bred modes’ for generating initial perturbations used in ensemble prediction. *Tellus* **51A**, 431–441.
- Evensen, G. **1994**. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res.* **99**(C5), 10143–10162.
- Evensen, G. **2003**. The ensemble Kalman filter: theoretical formulation and practical implementation. *Ocean Dynam.* **53**, 343–367.
- Feng, J., Ding, R. Q., Liu, D. Q. and Li, J. P. **2014**. The application of nonlinear local Lyapunov vectors to ensemble predictions in Lorenz systems. *J. Atmos. Sci.* **71**(9), 3554–3567.
- Fisher, M. **1996**. The specification of background error variances in the ECMWF variational analysis system. In: *Proceedings ECMWF Seminar on Data Assimilation*, Reading, UK, pp. 645–652. Available from ECMWF.
- Fisher, M. **2003**. Background error covariance modelling. In *ECMWF Seminar on Recent Developments in Data Assimilation for Atmosphere and Ocean*, Reading UK, 8–12 September. ECMWF.
- Hamill, T. M. and Whitaker, J. S. **2011**. What constrains spread growth in forecasts initialized from ensemble Kalman filters? *Mon. Wea. Rev.* **139**, 117–131.
- Houtekamer, P. L., Lefaiivre, L. and Derome, J. **1995**. The RPN ensemble prediction system. In: *Seminar Proceedings on Predictability*, Vol. II, Reading, UK, European Centre for Medium-Range Weather Forecasts, pp. 121–146.
- Houtekamer, P. L. and Mitchell, H. L. **1998**. Data assimilation using an ensemble Kalman filter technique. *Mon. Wea. Rev.* **126**, 796–811.
- Houtekamer, P. L., Mitchell, H. L., Pellerin, G., Buehner, M., Charron, M. and co-authors **2005**. Atmospheric data assimilation with an ensemble Kalman filter: results with real observations. *Mon. Wea. Rev.* **133**, 604–620.
- Kalnay, E. **2002**. *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge University Press, New York, NY, 211 pp.
- Lagarias, J. L., Reeds, J. A., Wrights, M. H. and Wright, P. E. **1998**. Convergence properties of the Nelder-Mead simplex method in low dimensions. *SIAM J. Optim.* **9**, 112147.
- Li, J. P. and Ding, R. Q. **2011**. Temporal–spatial distribution of atmospheric predictability limit by local dynamical analogs. *Mon. Wea. Rev.* **139**, 3265–3283.
- Lorenz, E. N. **1963**. Deterministic nonperiodic flow. *J. Atmos. Sci.* **20**, 130–141.
- Lorenz, E. N. **1965**. A study of the predictability of a 28-variable atmospheric model. *Tellus* **17**, 321–333.
- Lorenz, E. N. **1982**. Atmospheric predictability experiments with a large numerical model. *Tellus* **34**, 505–513.
- Marshall, J. and Molteni, F. **1993**. Toward a dynamical understanding of planetary-scale flow regimes. *J. Atmos. Sci.* **50**(12), 1792–1818.
- Miyoshi, T. and Kadowaki, T. **2008**. Accounting for flow-dependence in the background error variance within the JMA global four-dimensional variational data assimilation system. *SOLA* **4**, 37–40.
- Molteni, F., Buizza, R., Palmer, T. N. and Petroliagis, T. **1996**. The ECMWF ensemble prediction system: methodology and validation. *Quart. J. Roy. Meteor. Soc.* **122**, 73–119.
- Parrish, D. F. and Derber, J. C. **1992**. The National Meteorological Center’s spectral statistical-interpolation analysis system. *Mon. Wea. Rev.* **120**, 1747–1763.
- Peña, M. and Toth, Z. **2014**. Estimation of analysis and forecast error variances. *Tellus* **66A**, 21767.
- Privé, N. C., Errico, R. M. and Tai, K.-S. **2013**. The influence of observation errors on analysis error and forecast skill investigated with an observing system simulation experiment. *J. Geophys. Res.* **118**, 5332–5346.
- Ravela, S. **2007**. Two new directions in data assimilation by field alignment. *Lecture notes in Computer Science, Proc. ICCS* **4487**, 1147–1154.
- Ravela, S., Emanuel, K. and McLaughlin, D. **2007**. Data assimilation by field alignment. *Phys. D* **230**, 127145.
- Toth, Z. and Kalnay, E. **1993**. Ensemble forecasting at NMC: the generation of perturbations. *Bull. Amer. Meteor. Soc.* **74**, 2317–2330.
- Toth, Z. and Kalnay, E. **1997**. Ensemble forecasting at NCEP and the breeding method. *Mon. Wea. Rev.* **125**, 3297–3319.
- Trevisan, A. and Legnani, R. **1995**. Transient error growth and local predictability: a study in the Lorenz system. *Tellus* **47A**, 103–117.
- Wei, M., Toth, Z., Wobus, R. and Zhu, Y. **2008**. Initial perturbations based on the ensemble transform (ET) technique in the NCEP global operational forecast system. *Tellus* **60A**, 62–79.
- Wei, M., Toth, Z. and Zhu, Y. **2010**. Analysis differences and error variance estimates from multi-centre analysis data. *Aust. Meteorol. Ocean J.* **59**, 25–34.
- Wang, H., Huang, X., Sun, J., Xu, D., Zhang, M., Fan, S. and Zhong, J. **2014**. Inhomogeneous background error modeling for WRF-Var using the NMC method. *J. Appl. Meteor. Climatol.* **53**(10), 2287–2309.
- Whitaker, J. S., Hamill, T. M., Wei, X., Song, Y. and Toth, Z. **2008**. Ensemble data assimilation with the NCEP global forecast system. *Mon. Wea. Rev.* **136**, 463–482.

Appendix A. The EnKF and its application

True states denoted by $\mathbf{x}^t$ are generated by integrating the QG model for a long time period. Simulated observations $\mathbf{y}$ are produced from the true state using:

$$\mathbf{y} = \mathbf{H}\mathbf{x}^t + \boldsymbol{\varepsilon}, \quad (\text{A1})$$

where $\mathbf{H}$ provides a mapping from the model space to the observation space, and $\boldsymbol{\varepsilon}$ are the observation errors of GH (m²). Following Houtekamer and Mitchell (1998), $\boldsymbol{\varepsilon}$ are from the observational error covariance matrix $\mathbf{R}$:

$$\mathbf{R} = \overline{(\boldsymbol{\varepsilon}_{200}, \boldsymbol{\varepsilon}_{500}, \boldsymbol{\varepsilon}_{800})^T (\boldsymbol{\varepsilon}_{200}, \boldsymbol{\varepsilon}_{500}, \boldsymbol{\varepsilon}_{800})} = \begin{pmatrix} 260 & 0 & 0 \\ 0 & 80 & 0 \\ 0 & 0 & 30 \end{pmatrix}. \quad (\text{A2})$$

Observations are provided every 12 h. The ensemble of forecast states is obtained as background states. Initial background states are generated by superposing random errors from the covariance matrix $\begin{pmatrix} 390 & 80 & 20 \\ 80 & 120 & 20 \\ 20 & 20 & 50 \end{pmatrix}$ to true states. The ensemble background matrix

$\mathbf{X}^f$ for a given observation time is defined by the ensemble background forecasts $\mathbf{x}_i^f$ as:

$$\mathbf{X}^f = (\mathbf{x}_1^f, \mathbf{x}_2^f, \dots, \mathbf{x}_N^f), \quad (\text{A3})$$

where $N = 200$ here. The ensemble perturbation matrix can then be expressed as:

$$\mathbf{X}^{f'} = (\mathbf{x}_1^f - \overline{\mathbf{x}^f}, \mathbf{x}_2^f - \overline{\mathbf{x}^f}, \dots, \mathbf{x}_N^f - \overline{\mathbf{x}^f}) \quad (\text{A4})$$

where $\overline{\mathbf{x}^f}$ denotes the mean of the ensemble. The covariance matrix of the ensemble $\mathbf{X}^f$ is:

$$\mathbf{P}^f = \frac{1}{N-1} \mathbf{X}^{f'} \mathbf{X}^{f'T}. \quad (\text{A5})$$

The background ensemble is updated by a set of perturbed observation vectors $\mathbf{y}_i$ ($i = 1, 2, \dots, N$):

$$\mathbf{y}_i = \mathbf{y} + \boldsymbol{\varepsilon}_i. \quad (\text{A6})$$

Assuming perfect knowledge of the observational error variance, the random perturbations $\boldsymbol{\varepsilon}_i$ follow the same covariance matrix as $\boldsymbol{\varepsilon}$. The observations are assimilated to produce a new analysis of the state:

$$\mathbf{x}_i^a = \mathbf{x}_i^f + \mathbf{K}(\mathbf{y}_i - \mathbf{H}\mathbf{x}_i^f). \quad (\text{A7})$$

The Kalman gain $\mathbf{K}$ is calculated by:

$$\mathbf{K} = \mathbf{P}^f \mathbf{H}^T (\mathbf{H} \mathbf{P}^f \mathbf{H}^T + \mathbf{R})^{-1}. \quad (\text{A8})$$

$\mathbf{K}$ is actually a weighting measuring the ratio of the forecast and observational error covariance which determines to what extent the background forecasts will be corrected to fit the observations.

To avoid the collapse of the ensemble (i.e. filter divergence, Houtekamer and Mitchell, 1998) due to EnKF approximations including the undersampling of the full phase space of the MM-QG system (6144 total degrees of freedom) with only 200 perturbations, the magnitude of first guess perturbations are inflated by 30%.⁴ No background covariance localisation scheme is applied since the use of 200 ensemble members reduces the presence of spurious long-distance correlations. The assimilation cycle is repeated every 12 h for a 90-day period and initial conditions for single (unperturbed) forecasts are created by taking the mean of the analysis ensemble $\mathbf{x}_i^a$ ($i = 1, 2, \dots, N$).

Appendix B.

The NMC method

A common method to estimate the background error covariance matrix widely used in operation is the NMC method (Parrish and Derber, 1992). The method uses differences between pairs of forecasts with different lead times, valid at the same time (i.e. lagged forecast differences). In global scale applications, the background error covariance $\mathbf{B}$ is the statistics of differences between 48- and 24-h forecasts aggregated over time and/or space:

$$\mathbf{B} = \overline{(\mathbf{x}^{48} - \mathbf{x}^{24})(\mathbf{x}^{48} - \mathbf{x}^{24})^T} \quad (\text{B1})$$

where $\mathbf{x}^{48}$ and $\mathbf{x}^{24}$ are 48- and 24-h forecasts, respectively, valid at the same time. In regional-scale applications, usually 24- and 12-h lagged forecasts are used in place of 48- and 24-h forecasts. The spatial distribution of $\mathbf{B}$ is often smoothed and the magnitude of the matrix can also be tuned for optimal performance in DA schemes.