

Using Observed Data for Testing the Statistical Consistency of Initial Ensemble Perturbations

István Szunyogh¹

University of Maryland, IPST and Department of Meteorology

Györgyi Gyarmati,

Eötvös Loránd University, Department of Meteorology, Budapest, Hungary

Dezső Dévényi

NOAA, Forecast Systems Laboratory and University of Colorado, CIRES

January 15, 2003

¹*Corresponding author:* University of Maryland, Institute for Physical Science and Technology, Computer and Space Sciences Building, College Park, MD, 20742-2431, E-mail:szunyogh@ipst.umd.edu

Abstract

Two random variables are defined, one by an initial ensemble perturbation at a given geographical location, and another one by the analysis uncertainty at the same location. The initial ensemble perturbations are said to be statistically consistent with the analysis uncertainty if the probability density functions of the two aforementioned random variables are equal. In this paper, it is shown how observed data can be used to test this statistical consistency.

The usefulness of the proposed approach is demonstrated by an application to the global ensemble forecasting system of the National Centers for Environmental Prediction. Targeted dropsonde observations, collected during the 2000 Winter Reconnaissance program, are used to show that increasing the rescaling frequency in the breeding cycle, from once a day to four times a day, improves the consistency of the initial ensemble perturbations.

1 Introduction

The performance of an ensemble prediction system, especially for short forecast lead times, is mainly determined by the representativeness of the initial ensemble perturbations for the analysis uncertainty. It is important to verify, therefore, that the initial ensemble perturbations correctly represent the analysis uncertainty. The chief difficulty with this verification is that the true state of the atmosphere is unknown and the analysis uncertainty (the expected difference between analysis and truth) cannot be directly determined.

The most frequently used technique to avoid the aforementioned problem is to carry out Observing System Simulation Experiments (OSSE; Houtekamer 1995; Hamill et al. 2000). In these experiments the "true state" of the atmosphere is obtained by a model integration, hence the analysis error statistics can be directly determined from a sample of the difference between the analysis and the "true state". While the OSSEs have provided important insights into the strengths and the weaknesses of the different ensemble generation techniques, they cannot be used to measure the performance of an operational ensemble prediction system.

In the operational practice, the quality of the initial perturbations is usually measured by the performance of the ensemble at short (6-72-hour) forecast lead times, which is inferred by replacing the true state by its best estimate, the analysis. There are two important limitations of this approach. Firstly, the analysis errors are not negligible compared to the forecast errors, especially for the very short lead times. Secondly, forecast errors are due not only to growing analysis but also to model errors. The performance of the ensemble perturbation generation technique, therefore, cannot be separated from the short term performance of the forecast model.

In this paper, we propose an alternative verification technique, for which the results are not affected by model errors. Our technique exploits the opportunity provided by a unique set of targeted dropsonde observations collected during the 2000 Winter Storm Reconnaissance (WSR00) field program over the northeast Pacific. This data set has numerous favorable features (Szunyogh et al., 2002):

- The instrument errors are small for the University Corporation for Atmospheric Research Global Positioning System (UCAR GPS) dropsondes that collected the data.
- The data were collected in regions of significant synoptical features that later had significant impact on the weather over the United States.
- Once assimilated, the data led to significant improvements in the 1-5 days forecasts over the United States, which indicates their capability of detecting errors in the analysis cycle that did not assimilate them.

Section 2 explains how these data can be used to validate ensembles generated by perturbing analyses that were prepared by excluding the targeted dropsonde observations. Section 3 demonstrates the usefulness of the technique by an application to two sets of bred perturbations. One set was obtained from the operationally used 24-hour breeding cycle, while the other was generated by a 6-hour breeding cycle.

2 Verification technique

Let y^t , y^o , y^e , and x^a be random variables that represent, respectively, the true state, the observed state, the observational error, and the analysis at the observational location. Then,

$$(y^o - x^a) = (y^t + y^e) - x^a = (y^t - x^a) + y^e. \quad (1)$$

Our goal is to test whether the initial ensemble perturbations are representative for the difference between the true state and the analysis, $(y^t - x^a)$. This goal is achieved by (i) first replacing $(y^t - x^a)$ by the random variable x^p that represents an initial ensemble perturbation at the observational location and then (ii) comparing the two sides of (1). Since the variables are random, the equality of $(y^o - x^a)$ and $x^p + y^e$ can be verified only in a statistical sense.

Suppose that (i) sufficiently large samples of $(y^o - x^a)$ and x^p are available to estimate their probability density functions (pdf.s), $f(z)$ and $h(z)$, respectively; and (ii) an estimate of the observational error, $g(z)$, is available. Our goal is to verify using this information whether

the initial ensemble perturbations are *statistically consistent* with the analysis uncertainty, i.e. whether $h(z)$ can be equal to the probability density function of $(y^t - x^a)$.

The random variables that represent the analysis uncertainty, $(y^t - x^a)$, and the observational error, y^e , become independent, if the observations used to compute $(y^o - x^a)$ are not used in the analysis. This means that the initial ensemble perturbation represents a random variable which is independent of the observational error. Therefore, the initial ensemble perturbation can be consistent with the analysis uncertainty only if the pdf. of $(y^o - x^a)$ is equal to the pdf. of $x^p + y^e$; i.e.

$$f(z) = \int_{-\infty}^{+\infty} h(z - v)g(v)dv. \quad (2)$$

In summary, the statistical consistency between the ensemble perturbation and the analysis uncertainty can be verified by the following algorithm:

- Compute the empirical density functions $f(z)$ and $h(z)$ based on a sample of $(y^o - x^a)$ at the observational locations and a sample of ensemble perturbations, x^p , at the same observational locations.
- Choose a probability density function, $g(z)$, to represent the distribution of the observational errors.
- Compute the convolution $c(z)$ of $h(z)$ and $g(z)$:

$$c(z) = \int_{-\infty}^{+\infty} h(z - v)g(v)dv, \quad (3)$$

- Compare $c(z)$ and $h(z)$.

We note that one might (falsely) assume that there exist an alternative verification technique, in which (2) is applied to the alternative

$$(y^t - x^a) = (y^o - x^a) - y^e. \quad (4)$$

form of (1). This way the pdf. of $(y^t - x^a)$ could be directly determined from $f(z)$ and $g(z)$ and compared to $h(z)$. This approach does not work, however, because the two terms on the

rhs. of (4) are not independent random variables; the difference between the observation and the analysis is dependent on the observational error.

3 Application

During the WSR00 field program, nearly 300 dropsondes were released over the northeast Pacific on 12 separate flight days. To assess the forecast effect of the data, an additional analysis-forecast cycle was run parallel to the operational cycle. This parallel cycle was identical to the operational one, except no dropsonde data were assimilated (see Szunyogh et al. 2002 for more details). The ensemble initial conditions evaluated in this study were prepared by perturbing analyses from the parallel cycle. This was done to ensure that the difference between the dropsonde observations and the unperturbed analyses is independent of the observational errors.

The computation of the difference between the observation and the analysis at the observational location required the interpolation of the analysis field, which was originally available on the Gaussian grid of the spectral NCEP model. The same interpolation was also needed to obtain the initial ensemble perturbations at the observational locations. Statistical samples were collected for the horizontal wind components and the virtual temperature at the 700 hPa, 500 hPa, and 300 hPa pressure level and for the surface pressure. At these levels observations were available from each sonde dropped during the field program.

The most difficult task is to find a proper statistical model for the observational errors. In this paper, the observational errors are assumed to be normally distributed and results will be shown for two different choices of the standard deviation. One of them is the manufacturer provided instrument error for the NCAR GPS sondes (Hock and Franklin 1999) while the other one is the value assumed by the operational NCEP data assimilation system. The former choice is expected to provide an estimate of the lower bound of the observational error since, beyond the instrument error the observational error also has an elusive representativeness error component. The observational errors assumed by the operational analysis

scheme, on the other hand, include an implicit estimate of the representativeness error, since they are experimentally tuned to provide optimal forecast performance. The problem with this estimate is that the tunable parameter, the assumed observational error, may also be affected by other imperfectness of the analysis scheme (e.g. problems with the background error covariance matrix).

Two sets of initial ensemble perturbations were generated by running breeding cycles (Toth and Kalnay, 1997) for the duration of the WSR00 field program. In one of these cycles, the initial perturbations were generated by that version of the breeding algorithm, which was in operational use in 2000. The other cycle was identical, except for that the perturbations were rescaled every six instead of every twenty-four hours. This change was expected to improve the consistency of the initial ensemble perturbations with the analysis uncertainty. In what follows, our verification technique explained in section 2, is used to investigate whether this expectation was fulfilled by the experimental ensemble.

For the surface pressure, the results presented in Figure 1 suggest the following conclusions: (1) the frequency of the large analysis errors is overestimated, while the frequency of the small analysis errors is underestimated by the ensembles (2) this problem is more serious for the ensemble with 24-hour rescaling frequency (3) the performance of the ensemble is found to be better when the smaller estimate of the observational error (instrumental error) is used in the verification.

Interestingly, the above conclusions remain valid regardless of which state variable (temperature or wind) is used, even though the pdf.s show a strong dependence on the variable (Figure 2-Figure 4). The good agreement between the results for the different state variables indicates a systematic problem with the initial perturbations. To further investigate this problem, it is useful to define a measure that can characterize the difference between the random variables compared by one single number. A straightforward choice is

$$\left(\frac{\sqrt{\langle x^p{}^2 \rangle}}{\sqrt{\langle (y^o - x^a)^2 \rangle - \langle y^{e2} \rangle}} - 1 \right) \times 100, \quad (5)$$

which shows, in percentages, the extent to which the root-mean-square (rms) of the analysis uncertainty is overestimated by the initial ensemble perturbations. (The angled bracket denotes sample mean). The results summarized in Table 1 corroborate the conclusions drawn by visually inspecting the estimated probability density functions: The ensemble perturbations typically overestimate the analysis uncertainty. The only exception is the wind at 300hPa pressure level when the smaller estimate of the observational error is considered. As it can be expected by considering the specific form of Equation 5, the overestimation of the analysis uncertainty is always found to be a less serious problem when the smaller estimate of the observational error is used. What cannot be expected by considering only the equations is that the overestimation is always smaller for the more frequent rescaling. Nevertheless, this result is in good accordance with the conclusions drawn based on comparing the pdf.s. If the grossly overestimated surface pressure is not counted and it is assumed that the values obtained with the two different estimates of the observational error can be averaged, it is found the initial ensemble perturbations overestimate the analysis uncertainty, on average, by 38,7% in the 6-hour breeding cycle, and by 67.1 % in the 24-hour cycle. We note, that it is not surprising to find some overestimation of the analysis uncertainty. The initial size of the ensemble perturbation is tuned that way, that the ensemble spread (defined by the standard deviation in the ensemble) is equal to the average error in the ensemble mean at three days forecast lead time. Since the error in the ensemble mean grows faster than the ensemble spread, the initial magnitude of the ensemble perturbations must be somewhat overestimated. The large scale area mean of the magnitude of the individual ensemble perturbation is constant by design (equal for the 6 and 24 hour cycles). Our result, therefore, show that in the regions of the atmospheric instabilities, where the observations were collected, the 6 hour rescaling provides perturbations, which are more consistent with the analysis uncertainty.

The relatively large variability of the numbers for the different state variables in Table 1 can be attributed to a couple of possible factors. Firstly, the representativeness error can be different for the different variables. Secondly, at a given geographical location the breeding

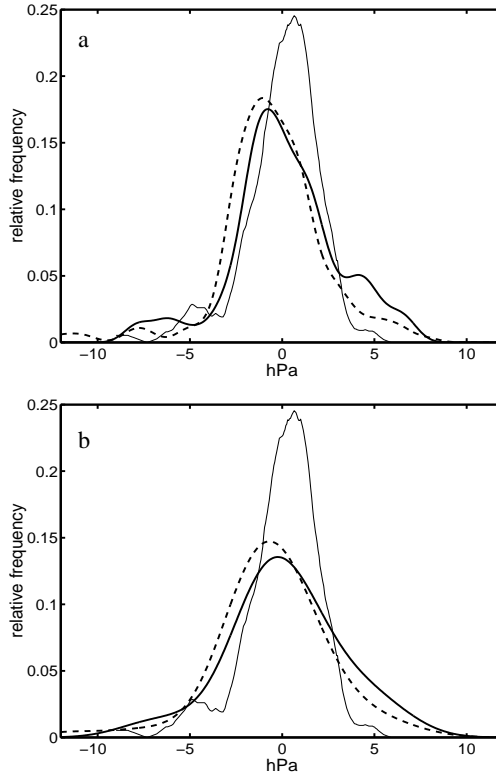


Figure 1: Probability density functions for the surface pressure. Shown are the pdf.s for $y^o - a^x$ (thin solid line), $x^p + y^e$ with 24-hour rescaling (solid line), and $x^p + y^e$ with 6-hour rescaling (dashes). The manufacturer provided instrument error is $y^e = 0.5 \text{ hPa}$ (panel a), while the observational error assumed by the data assimilation scheme is $y^e = 1.6 \text{ hPa}$ (panel b).

algorithm rescales all variables with the same factor, assuming that there is a linear balance between the variables. While this is a reasonable assumption at the synoptic and the larger scales, there may be strong local deviations from this rule in a model based on the primitive equations.

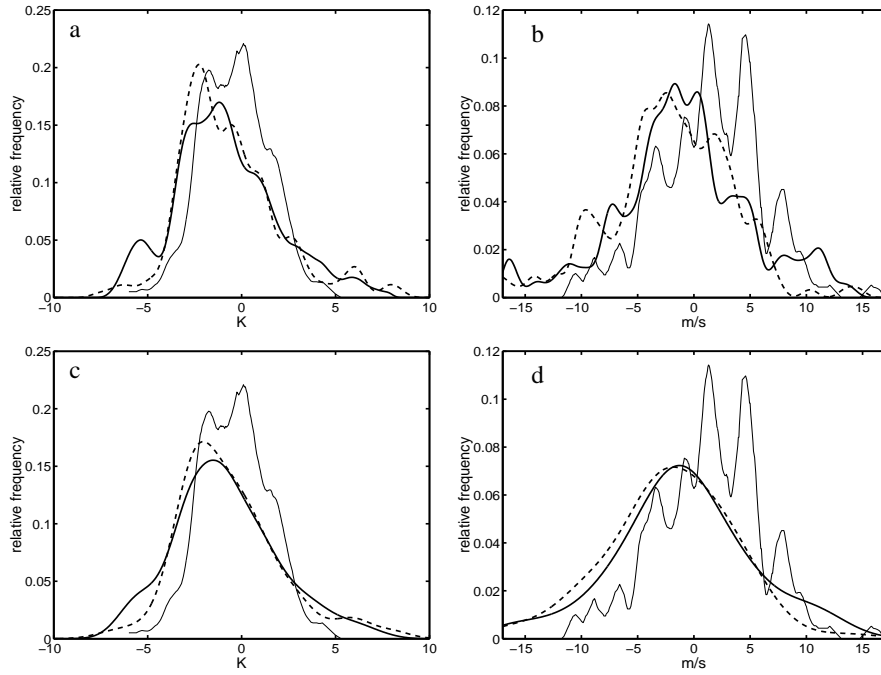


Figure 2: Probability density functions for the temperature and wind at 700hPa. Shown are the pdf.s for $y^o - a^x$ (thin solid line), $x^p + y^e$ with 24-hour rescaling (thick solid line), and $x^p + y^e$ with 6-hour rescaling (dashes). The manufacturer provided instrument error for temperature is $y^e = 0.2 K$ (panel a), while the observational error assumed by the data assimilation scheme is $y^e = 0.8 K$ (panel c). Panel b and d are the same as a and c, respectively, for the wind. The instrument error is 0.5 m/s, and the observational error is 2.4 m/s.

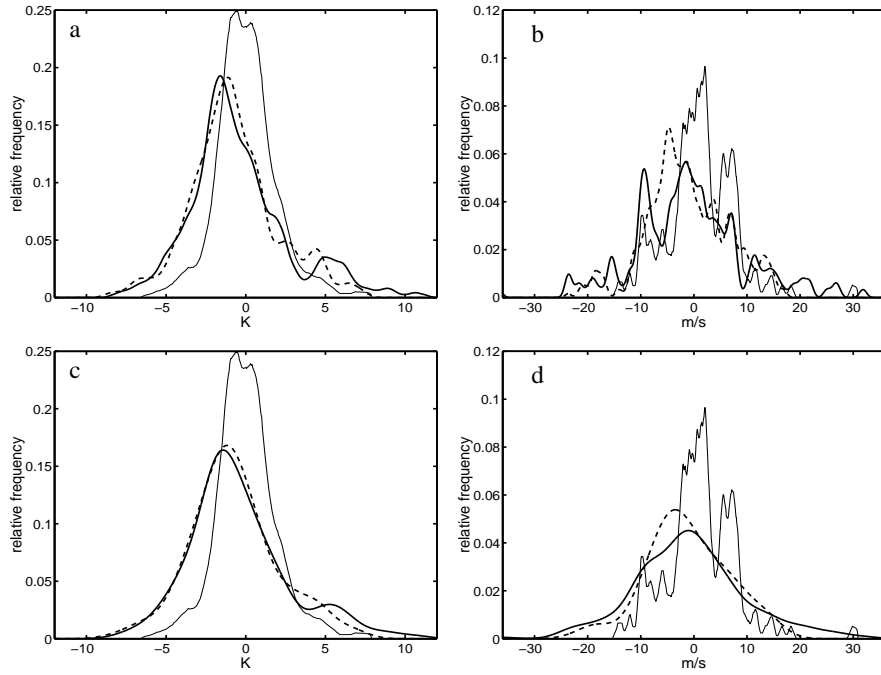


Figure 3: Probability density functions for the temperature and wind at 500hPa. Shown are the pdf.s for $y^o - a^x$ (thin solid line), $x^p + y^e$ with 24-hour rescaling (thick solid line), and $x^p + y^e$ with 6-hour rescaling (dashes). The manufacturer provided instrument error for temperature is $y^e = 0.2 K$ (panel a), while the observational error assumed by the data assimilation scheme is $y^e = 0.8 K$ (panel c). Panel b and d are the same as a and c, respectively, for the wind. The instrument error is 0.5 m/s, and the observational error is 2.8 m/s.

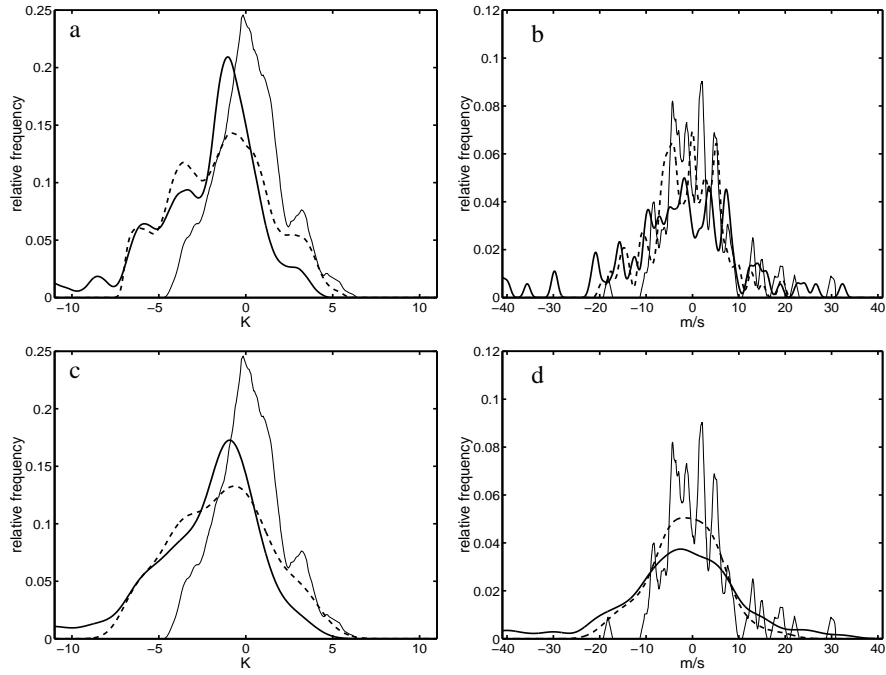


Figure 4: Probability density functions for the temperature and wind at 300hPa. Shown are the pdf.s for $y^o - a^x$ (thin solid line), $x^p + y^e$ with 24-hour rescaling (thick solid line), and $x^p + y^e$ with 6-hour rescaling (dashes). The manufacturer provided instrument error for temperature is $y^e = 0.2 K$ (panel a), while the observational error assumed by the data assimilation scheme is $y^e = 0.8 K$ (panel c). Panel b and d are the same as a and c, respectively, for the wind. The instrument error is 0.5 m/s, and the observational error is 2.8 m/s.

4 Concluding remarks

In this paper, a new and simple statistical technique was introduced to test the statistical consistency between initial ensemble perturbations and analysis uncertainties. Our analysis indicates that better statistical consistency can be achieved with the breeding technique when the ensemble perturbations are rescaled every 6 instead of every 24 hours. This result is not totally surprising, considering that the purpose of the breeding technique is to mimic the effects of the analysis cycle on the growing analysis errors. More precisely, it assumes that when observed data are assimilated the magnitude of the growing errors is reduced. Based on this consideration, the rescaling was done every 6 hours at the time (in 1992) when ensemble prediction was introduced into operation at NCEP. This meant, however, that an extra day of ensemble integration was needed to evolve the initial ensemble perturbations and a 24-hour rescaling was introduced soon to save computer time. In the new setup, the regular 24-hour ensemble forecasts were rescaled.

Partly based on the results presented here, NCEP is considering to return to the 6-hour rescaling later this year. The preliminary forecast verification results (not shown in this paper) indicate that increasing the rescaling frequency also improves the skill of the individual ensemble members.

Acknowledgments

The research presented in this paper was supported by the Keck Foundations.

References

- Hock, T. F. and Franklin, J. L.,1999: The NCAR GPS Dropwindsonde. *Bull.Am.Met.Soc.*,**80**, 407-420.
- Hamill, T. M., and C. Snyder, 2000: A hybrid ensemble Kalman filter-3D variational analysis

scheme. *Mon. Wea. Rev.*, **128**, 2905-2919.

Szunyogh, I., Toth, Z., Majumdar, S. J., and Persson, A., 2002: On the propagation of the effect of targeted observations: The 2000 Winter Storm Reconnaissance Program. *Monthly Weather Review*, **130**, (in print, accepted October 2001).

Toth, Z. and E. Kalnay, 1997: Ensemble forecasting at NCEP and the breeding method. *Mon. Wea. Rev.*, **125**, 3297-3319.

Variable	Level	y^e	06h rms %	24h rms %
pressure	surface	1.6	140	160
temperature	300	0.8	60	85
wind	300	2.8	6	84
temperature	500	0.8	68	93
wind	500	2.8	35	81
temperature	700	0.8	68	68
wind	700	2.4	32	45
pressure	surface	0.5	49	61
temperature	300	0.2	46	69
wind	300	0.5	-1	71
temperature	500	0.2	52	75
wind	500	0.5	22	63
temperature	700	0.2	52	52
wind	700	0.5	15	26

Table 1: Overestimation of the analysis uncertainty by the initial ensemble perturbations (for definition see Eq. 5). Shown are the selected state variables (first column), the pressure levels at which the observations were taken (second column), the estimated observational error (third column), overestimates in the 6-hour rescaling cycle (fourth column), and in the 24-hour cycle (fifth column)