IL NUOVO CIMENTO Vol. 31 C, N. 2 Marzo-Aprile 2008
DOI 10.1393/ncc/i2008-10288-2

A note on the multimodel superensemble technique for reducing forecast errors

L. Kantha([1]), S. Carniel([2])(*) and M. Sclavo([2])

([1]) *Department of Aerospace Engineering Sciences, 431 UCB, University of Colorado - Boulder CO 80309, USA*
([2]) *CNR, Instituto di Scienze Marine - Castello 1364, I-30122, Venice, Italy*

(ricevuto il 12 Febbraio 2008; approvato il 18 Giugno 2008; pubblicato online il 23 Settembre 2008)

Summary. — The multimodel superensemble (SE) technique has been used with considerable success to improve meteorological forecasts and is now being applied to ocean models. Although the technique has been shown to produce deterministic forecasts that can be superior to the individual models in the ensemble or a simple multimodel ensemble forecast, there is a clear need to understand its strengths and limitations. This paper is an attempt to do so in simple, easily understood contexts. The results demonstrate that the SE forecast is almost always better than the simple ensemble forecast, the degree of improvement depending on the properties of the models in the ensemble. However, the skill of the SE forecast with respect to the true forecast depends on a number of factors, principal among which is the skill of the models in the ensemble. As can be expected, if the ensemble consists of models with poor skill, the SE forecast will also be poor, although better than the ensemble forecast. On the other hand, the inclusion of even a single skillful model in the ensemble increases the forecast skill significantly.

PACS 07.05.Tp – Computer modeling and simulation.
PACS 92.10.-c – Physical oceanography.
PACS 95.75.-z – Observation and data reduction techniques; computer modeling and simulation.
PACS 95.75.Pq – Mathematical procedures and computer techniques.

1. – Introduction

Ever since they started producing weather forecasts for general consumption by the public, meteorologists have faced the problem of improving them, initially concentrating on the model physics, data-assimilation techniques, and on the relative database.

(*) E-mail: sandro.carniel@ismar.cnr.it

However, over the years, it has become apparent that even with a skillful model and abundant data, sensitivity to initial conditions of the nonlinear systems would not allow high-fidelity *deterministic* forecasts, leading to the so-called ensemble techniques: with a forecast repeated many times and an *ensemble* of forecasts prepared, it became then possible to produce *probabilistic* forecasts, which are more natural for a chaotic system like the atmosphere. Ensemble forecasts are now routine in both weather and seasonal forecasts, both in the USA and Europe (*e.g.*, at the National Centers for Environmental Prediction, NCEP, and at the European Centre for Medium-Range Weather Forecast, ECMWF, respectively), because of the attractiveness of the concept and also because of the enormous computing power at the disposal of modern forecasters. However, simply averaging over several realizations of a single model has its shortcomings too, since no model or forecast technique is perfect and hence the model bias and error affect the resulting ensemble forecast.

Given the fact that there may exist several models and techniques, each with its own advantages and shortcomings, it becomes advantageous to include many different, presumably skillful, models and techniques in the ensemble (*multimodel ensemble*). In operational meteorology, this essentially means including forecasts from several operational centers in the ensemble, once again made possible by modern computing and communication technologies.

Meteorologists have realized that it is possible to go even beyond this and substantially improve the forecast skill by assessing the past performance of different models and techniques and using this information to assign appropriate weights to the individual forecasts in the ensemble. This technique is now known widely as the *multimodel superensemble* (SE) technique [1, 2]. Generally speaking, it involves comparing the performance of individual models/techniques in the ensemble to observational data over a selected period in the past and statistically assessing the individual model skill. This information is then used to assign appropriate weights to the forecasts of individual models in the ensemble to arrive at a better forecast. It appears that this technique is vastly superior to a simple multimodel ensemble, simply because prior information is used to assess model strengths and weaknesses, and this is then used to minimize forecast errors.

Naturally, the skill of the SE technique should depend on the length of the "training" period and the quality and the amount of data available for training. Very much akin to statistical forecasting, the skill depends on whether the data used for training depict accurately, *in the past*, the phenomena to be forecasted at present. The longer the training period, the better, since the training data are then likely to include more realizations similar to the events or features being forecasted. And it is here that one runs into data inadequacies and errors, especially in a field such as oceanography, where observational data tend to be sparse in time and space as well as duration [3, 4]. Nevertheless, by judicial choices, it may be possible to reduce the forecast errors significantly via the SE technique, as shown in recent years in meteorological forecasts [1, 5]. The technique has been applied to improve hurricane track and intensity forecasts (*e.g.*, [6]) and significant improvement in skill has been demonstrated. It has also been applied to seasonal forecasts (see [2]). The EU project DEMETER [7] was aimed at developing and validating a multi-model ensemble forecast system for reliable seasonal to interannual predictions. In the ocean forecasting arena, Rixen and Coelho [8] have applied the SE technique to operational ocean predictions, particularly to predictions of acoustic properties off the west coast of Portugal.

While the SE technique has been explored in the context of operational forecasts using complex atmosphere and ocean models, as mentioned above, or for seasonal

forecasts [1, 2], there is still a need to understand the strengths and limitations of this technique in the context of simple, easily understood examples. This paper is an attempt to do so in the case of deterministic short-term forecasts, appropriate to say weather forecasts rather than long-term forecasts such as seasonal forecasts.

In short, the SE technique enables individual model biases to be diagnosed and removed and the best combination of the ensemble to be obtained by appropriately weighting the original models in the ensemble according to their error signature. The general idea behind is then that of the minimization of a mean squared error with respect to observations, carried out through a linear regression following standard approaches (*e.g.*, [5,9]) that allows the computations of appropriate coefficients or weights (x).

The SE forecast is constructed from weighted, bias-removed multimodel results:

$$X_{\mathrm{SE}}(t) = X_{\mathrm{mean}} + \frac{\sum_{n=1}^{N} x_n \left[Y_n(t) - Y_n^{\mathrm{mean}}\right]}{\sum_{n=1}^{N} x_n} \qquad T_T < t \le (T_T + T_F), \tag{1.1}$$

where N is the number of models in the ensemble, Y_n^{mean} is the model bias determined by averaging $Y_n(t)$ over the training period (TP) and x_n are the weights assigned to individual models in the ensemble, based on their skills. By definition, observational forecast mean (true mean) is not available over the forecast period (FP), therefore X_{mean} must be suitably selected.

The simple ensemble forecast is derived from a "blind" averaging of the ensemble:

$$X_{\mathrm{E}}(t) = \frac{1}{N} \sum_{n=1}^{N} Y_n(t) \qquad T_T < t \le (T_T + T_F). \tag{1.2}$$

If we define the root mean squares and the correlation for the SE cases as

$$RMS_{\mathrm{SE}} = \sqrt{\frac{1}{T_F} \int_{t=T_T}^{T_T+T_F} \left\{ \left[X_{\mathrm{SE}}(t) - X_{\mathrm{SE}}^{\mathrm{mean}}\right] - \left[X(t) - X_{\mathrm{mean}}\right] \right\}^2 \mathrm{d}t}, \tag{1.3}$$

$$COR_{\mathrm{SE}} = \frac{\int_{t=T_T}^{T_T+T_F} \left[X_{\mathrm{SE}}(t) - X_{\mathrm{SE}}^{\mathrm{mean}}\right] \left[X(t) - X_{\mathrm{mean}}\right] \mathrm{d}t}{\sqrt{\int_{t=T_T}^{T_T+T_F} \left[X_{\mathrm{SE}}(t) - X_{\mathrm{SE}}^{\mathrm{mean}}\right]^2 \mathrm{d}t} \sqrt{\int_{t=T_T}^{T_T+T_F} \left[X(t) - X_{\mathrm{mean}}\right]^2 \mathrm{d}t}} \tag{1.4}$$

with RMS_{E} and COR_{E} for the simple ensemble defined similarly, then the superensemble Forecast Skill Improvement (FSI) over that of a simple ensemble can be represented by the three indices defined as

$$FSI_{\mathrm{RMS}} = \frac{RMS_{\mathrm{E}}}{RMS_{\mathrm{SE}}} - 1, \quad FSI_{\mathrm{COR}} = \left| \frac{COR_{\mathrm{SE}}}{COR_{\mathrm{E}}} \right| - 1,$$
$$FSI_{\mathrm{BIAS}} = \frac{|X_{\mathrm{E}}^{\mathrm{mean}} - X_{\mathrm{mean}}|}{|X_{\mathrm{SE}}^{\mathrm{mean}} - X_{\mathrm{mean}}|} - 1, \tag{1.5}$$

where the numerator of the latter represents the bias for the simple multimodel ensemble cases ($BIAS_{\mathrm{E}}$) and the denominator that for the superensemble cases ($BIAS_{\mathrm{SE}}$). Values higher than zero for FSI_{RMS}, FSI_{COR} and FSI_{BIAS} indicate a higher skill for the SE vis-à-vis the regular ensemble. However, the absolute skill w.r.t. the true forecast is

indicated by RMS_{SE} and COR_{SE}([1]). Two possibilities exist for choosing X_{mean} in eq. (1.1). Traditionally, the observation mean (true mean) over the training period (which we will call *tmean*) has been used. However, another possibility is to simply use the ensemble mean of the N models in the ensemble over the forecast period (which we will call *emean*).

2. – Simpler first-order example

To illustrate the SE technique, Krishnamurti *et al.* [2] used the low-order Lorenz system, whose governing equations are

$$\begin{aligned} \frac{\mathrm{d}X}{\mathrm{d}t} &= -\sigma X + \sigma Y + f\cos\theta, \\ \frac{\mathrm{d}Y}{\mathrm{d}t} &= -XZ + rX - Y + f\sin\theta, \\ \frac{\mathrm{d}Z}{\mathrm{d}t} &= XY - bZ, \end{aligned} \tag{2.1}$$

where X, Y and Z are the three dependent variables of the third-order system. The constant σ is interpreted to be a dissipation/diffusion coefficient, r as a heating term and b the inverse of a scale height. The terms involving f are forcing terms. The constants σ, r and b denote the parameter space of the Lorenz attractor, with σ being the Prandtl number, r, the ratio of the Rayleigh number to its critical value, and b is the size of the convective cells; $f = 0$ in the original Lorenz attractor (1953).

In their study([2]), Krishnamurti *et al.* [2] set $\sigma = 50$, $r = 24.74$, $b = 13.35$, $f = 2.5$, and $\theta = 45°$ to define the base case, what they call the "Nature run". The parameter space defined by this particular choice of σ, r and b renders the system non-chaotic, unlike the original choice ($\sigma = 10$, $r = 28$, $b = 8/3$, $f = 0$) of Lorenz [10], which leads to a highly chaotic system. To derive the ensemble of models, Krishnamurti *et al.* [2] introduced random perturbations in the values of the parameters σ, r, b and θ (within the range ± 12.5, ± 3, ± 5, $\pm 2.5°$, respectively) but kept the value of f constant at 2.5. The initial conditions were also perturbed slightly and randomly (although the amplitudes are not specified in the paper) from the values chosen for the base case: $X = 0$, $Y = 10$ and $Z = 0$. The result was an ensemble of models that yielded a slowly decaying, oscillatory behavior for the dependent variables X, Y and Z, with considerable amplitude and phase differences between the various models in the ensemble. The task is then to arrive at a methodology able to combine the ensemble results to yield values for the dependent variables as close to the base case as possible.

Krishnamurti *et al.* [2] ran the different models for a time interval T of 200 units. During the training period (TP, 70 units), the multi-model time series of X, Y and Z were regressed against those of the base case to arrive at the weights (regarded as invariant) to be assigned to each model of the ensemble when performing the forecast. This simple and elegant example demonstrated the efficacy of the SE technique and paved the way for its application to meteorological forecasts. However, the example

([1]) Abbreviations used in the text: SE = superensemble; E = simple ensemble; VAR = variance; M = mean; COR = correlation coefficient; RMS = Root Mean Square; BIAS = bias; FSI = Forecast Skill Improvement; TP = Training Period; FP = Forecast Period.

([2]) Errors in their Table 1 and Eq. (1) have been corrected.

cited by Krishnamurti *et al.* [2] is that of a non-chaotic decaying oscillator; while the governing equations are nonlinear and coupled, and hence allow for the possibility of chaotic behavior, this feature is not a central part of their study.

Similar results can be demonstrated using a first-order non-chaotic decaying oscillator; for reasons of brevity, we will not present this in detail, but instead will summarize the findings. The decaying oscillator demonstrated that accounting for the skill of the models in the ensemble using prior information (SE forecast) almost always yields more skillful forecasts than just taking a blind average over the ensemble of models irrespective of their skill (ensemble forecast). However, the absolute skill of SE depends on the skill of the models in the ensemble, since to be skillful the ensemble should contain models with high correlation coefficients during the TP. However, this is not always true: if the models in the ensemble have high correlation coefficients during the TP, but much smaller ones during the FP, the skill information deduced from the *training period* is not valid during the *forecast period*, therefore leading to a not skillful SE. Only if the models in the ensemble behave similarly during both the TP *and* FP, then the past performance is indicative of the future behavior. The longer the relative *training period*, the better the SE skill. Inclusion of even a single skillful model in the ensemble may increase the SE forecast skill significantly, while random noise reduces it, with higher noises producing higher reductions.

3. – Lorenz system

The above conclusions, however, pertain to a non-chaotic system. What happens if the system is chaotic? This is a more interesting question that Krishnamurti *et al.* [2] did not tackle. To answer this question, we appeal to the Lorenz attractor, but instead of the parameter space employed by Krishnamurti *et al.* [2], which yields a decaying oscillator, we explored the chaotic regime by using

$$\sigma = 22.0, \qquad r = 35.0, \qquad b = 2.67, \qquad f = 2.0, \qquad \theta = 45.0. \tag{3.1}$$

The different parameters were randomly perturbed to generate an ensemble of 10 models differing from the base model. All the models were run for 20000 time steps with $\mathrm{d}t = 0.001$ (therefore for a total $t = 20$) and only the time series $X(t)$ was used for the analysis. The first half of the series (until $t = 10$) was discarded; the first half of the remaining time series ($t < 5$) was used for training, and the second half ($t > 5$) for the forecast. Note that a simple fourth-order Runge-Kutta method was used to solve the Lorenz set of three first-order, coupled nonlinear ordinary differential equations. It is well known that the Lorenz system possesses a strange attractor characterized by two "regimes", that is, two regions in phase space, which are recurrently visited by the orbits of the system.

The basic idea is then to explore situations where chaotic transitions between the two regimes make accurate predictions quite difficult and discern its impact on the skill of SE technique.

Figure 1a shows the last part of the training and the forecast performances in the period $t = 5.0$–6.6 for two typical cases, while fig. 1b presents the correlation coefficients between the 10 models used in the ensemble and the "truth" during the training (TP) and forecast periods (FP) for both cases. The corresponding statistics obtained using the set (3.1) are summarized in table I. The top panel of fig. 1a shows Case 1, in which the superensemble technique (dark blue line) is able to provide results that are in excellent agreement with the "true" forecast (red line). On the other hand, Case 2 (presented in

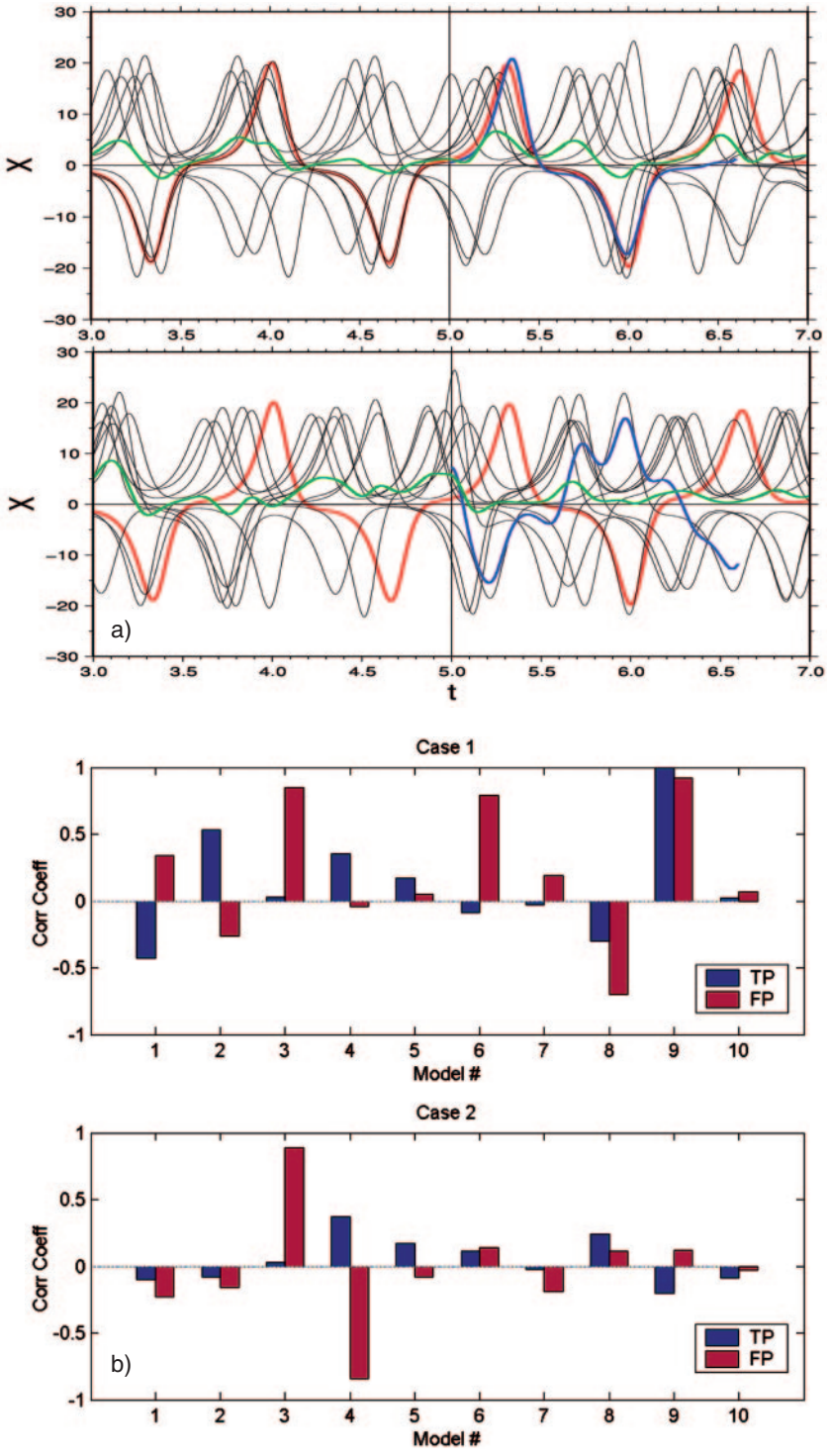

Fig. 1. – a) Time series of individual models (black lines), simple ensemble (green) and superensemble (dark blue) averages for Case 1 (top panel) and Case 2 (bottom panel). Red denotes the "truth", the vertical line the end of the TP (see table I for associated relevant statistics). b) The models in the ensemble and their correlation coefficients with "truth" for Case 1 and Case 2. TP: training period; FP: forecast period.

TABLE I. – *Statistics for Case 1 and Case 2. The "true" forecast mean (variance) is* 1.28 *(*78.67*).*

	N	VAR_{SE}	VAR_{E}	M_{SE}	M_{E}	COR_{SE}	COR_{E}	RMS_{SE}	RMS_{E}	$BIAS_{\mathrm{SE}}$	$BIAS_{\mathrm{E}}$	FSI_{COR}	FSI_{RMS}	FSI_{BIAS}
Case 1	10	64.68	5.42	−0.14	2.33	0.91	0.81	3.85	7.20	−1.42	1.05	0.13	0.87	−0.26
Case 2	10	75.89	2.01	−0.14	1.26	−0.81	−0.16	16.80	9.20	−1.42	−0.13	4.20	−0.45	−0.99

the bottom panel) is characterized by very poor results. The only difference between the two cases is the make-up of the models in the ensemble, as depicted by fig. 1b, the excellent performance in Case 1 is due to the inclusion of a highly skillful model (no. 9, with a correlation coefficient 1.0) in the ensemble. If we just use the results of model no. 9 for the forecast, the results would be equally good.

In Case 2, all the models in the ensemble had poor skill (maximum value for the correlation 0.37) and so the SE forecast reflects this. In both cases, the SE forecast performs clearly better than the simple ensemble forecast, as shown by the index FSI_{COR}. Note however that indices FSI_{RMS} and FSI_{BIAS} depend very much on the mean used in the forecast. Here the true training mean is used, and it happens to be slightly worse than the ensemble mean, consequently FSI_{BIAS} is negative in both cases. However, just looking at the FSI values can be misleading, since they only show the relative performance of the super-ensemble and ensemble averages; in Case 2, both are poor. What matters is the performance relative to the "true" forecast, as indicated by COR, RMS and $BIAS$. The closer the value of COR to 1.0, and the closer the values of RMS and $BIAS$ to zero, the better the forecast skill, *per se*. From this measure, SE performance is clearly better in Case 1 compared to Case 2.

In the above example the "true" system stays on the same regime of the attractor. It is interesting to explore what happens if the system stays on *one* regime during the training period, but transitions to the *other* during the forecast period. To examine this, we considered a slightly different parameter set for the Lorenz system

$$\sigma = 20.0, \qquad r = 28.0, \qquad b = 2.67, \qquad f = 2.0, \qquad \theta = 45.0. \tag{3.2}$$

Again, an ensemble of 10 models was generated by random perturbation of the parameters, and two typical results, called Case 3 and Case 4, are shown in fig. 2a, and the associated statistics in table II. Case 3 includes a skillful model (as determined from training, no. 9 with correlation coefficient 0.99), whereas Case 4 does not. From fig. 2b, showing again the correlations between the models and the truth, it is possible to note that the results are similar to the Cases 1 and 2 considered earlier, except for the overall poor performance of the SE technique. One important difference arises from the fact that the training mean (*tmean*) is much different from the true forecast mean; using *tmean* in the SE forecast leads to a large bias, and hence to worse values of FSI_{RMS} and FSI_{BIAS} (see values in parenthesis in table II). Slight improvements are seen when using the ensemble mean *emean*.

These examples suggest that the skill of the SE technique depends critically on the skill of the models in the ensemble, and not merely on the number of models adopted. This was confirmed by repeating the above runs with $N = 20$ and $N = 5$ (results not shown here).

However, the example considered above is somewhat extreme. A better approach, similar to that of Onken *et al.* [4], who used different hydrodynamical models employing different solution methodologies, methods of data assimilation and initializations, but nevertheless obtained consistent results, might be to take the basic Lorenz oscillator and perturb the *resulting time series*, introducing random modifications of the overall amplitude, phase and bias of the time series. An ensemble of 20 models was generated and used, producing two examples starting from the following sets:

$$\begin{aligned} &\text{Case L1:} \quad \sigma = 20.0, \quad r = 28.0, \quad b = 2.67, \quad f = 2.0, \quad \theta = 45.0, \\ &\text{Case L2:} \quad \sigma = 22.0, \quad r = 35.0, \quad b = 2.67, \quad f = 2.0, \quad \theta = 45.0. \end{aligned} \tag{3.3}$$

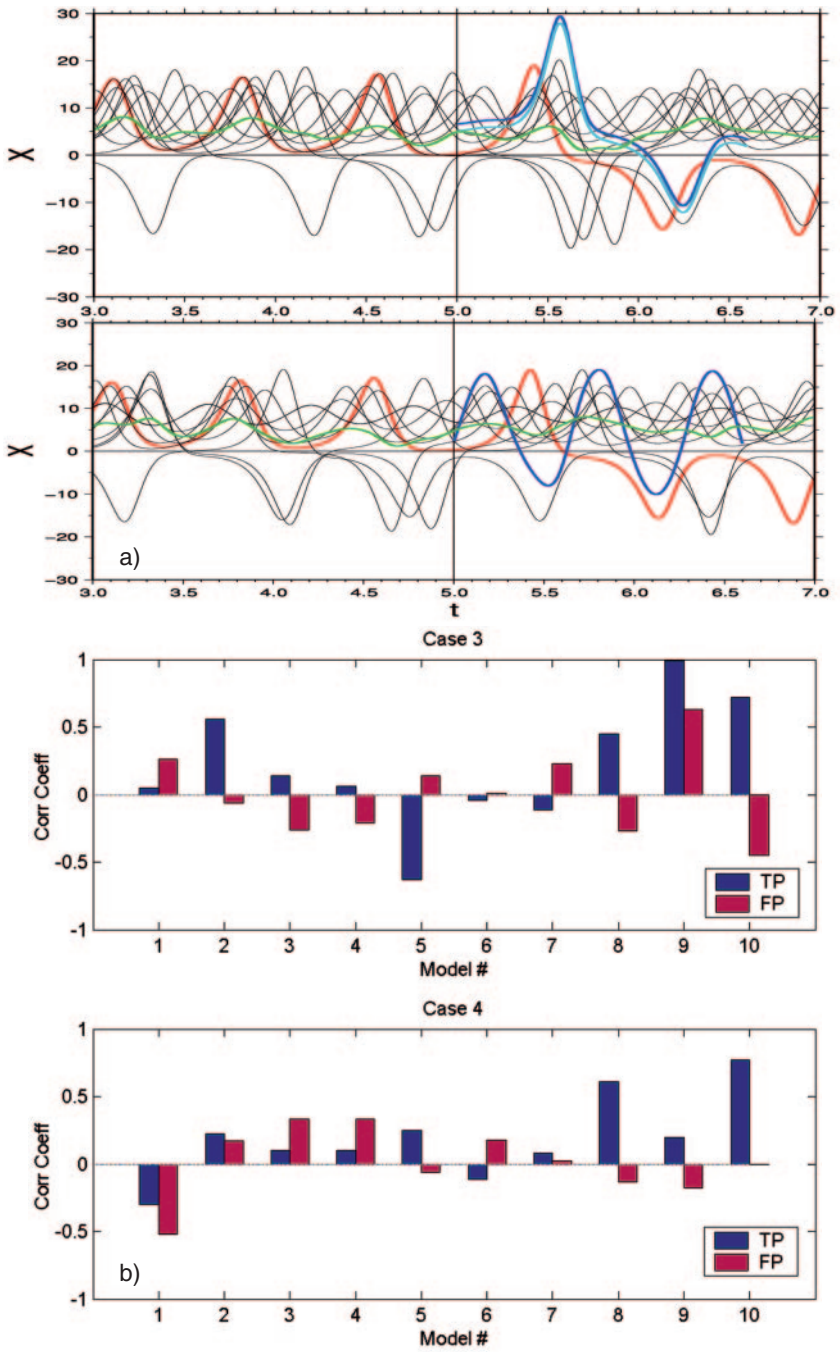

Fig. 2. – a) Time series of individual models (black lines), simple ensemble (green) and superensemble averages (dark blue: *tmean*, light blue: *emean*) for Case 3 (top panel) and Case 4 (bottom panel). Red denotes the "truth", the vertical line the end of the TP (see table II for associated relevant statistics). b) The models in the ensemble and their correlation coefficients with "truth" for Case 3 and Case 4. TP: training period; FP: forecast period.

TABLE II. – *Statistics for Case 3 and Case 4. The numbers in parentheses indicate values corresponding to the use of emean in the SE forecast. The "true" forecast mean (variance) is* -0.48 *(*59.41*).*

	N	VAR_{SE}	VAR_{E}	M_{SE}	M_{E}	COR_{SE}	COR_{E}	RMS_{SE}	RMS_{E}	$BIAS_{\mathrm{SE}}$	$BIAS_{\mathrm{E}}$	FSI_{COR}	FSI_{RMS}	FSI_{BIAS}
Case 3	10	81.74	3.38	5.81 (4.35)	4.35	0.61	0.04	9.76 (8.89)	9.22	6.30 (4.83)	4.83	14.59	-0.05 (0.04)	-0.23 (0.00)
Case 4	10	93.30	1.45	5.81	5.49	0.09	-0.11	13.40	9.92	6.30	5.97	-0.19	-0.26	-0.05

The training period lasts from $t = 2.4$ to $t = 5.0$, whereas the forecast period is from $t = 5.0$ to $t = 6.6$. The examples were chosen such that the training and forecast periods correspond to different regime of the attractor of the system in Case L1, and to the same regime in Case L2. The amplitude, phase and bias are varied randomly to yield the different models in the ensemble.

Figure 3a shows the results for both *tmean* and *emean*; in Case L1, *emean* just happens to be close to the true forecast mean and hence its use gives much better results compared to *tmean*. Indeed, since the true forecast mean (base case) differs significantly from the true training mean, its use produces a large bias in the forecast, which is also reflected in the RMS values (see Case L1 and Case L2 in table III). Figure 3b shows the models in the ensemble and the coefficients of correlation of each with the true time series during the training period.

On the other hand, in Case L2, the true forecast mean is not much different from the true training mean; consequently, the forecast results are essentially similar for *tmean* and *emean*. In any case, in both cases the SE is clearly more skillful than the simple ensemble, thus suggesting that the technique may turn out to be of great utility in practical applications to ocean (and atmosphere) forecasting.

To further explore the dependence of the performance of the SE method on the skill of the models in the ensemble, we ran additional cases starting from the sets (3.3). In Case 3-L1 and Case 3-L2, we retained only the models in the ensemble with correlation coefficients < 0.3 (w.r.t. the base case during the training period TP, see table III). This yielded an ensemble of 8 rather skill-less models. In Case 4-L1 and Case 4-L2, we retained all models with coefficients > 0.9, yielding two highly skillful models. These four results are shown in fig. 4 and the corresponding statistics in table III. It is clear that when the ensemble does not include skillful models, both the simple ensemble *and* the SE method fail to provide decent results, even though the SE technique is marginally better. The precise number of models included is not really that important, as can be seen from the cases when only two skillful models are included in the ensemble. In this case, even though only two models are present in the ensemble, the SE technique provides excellent results, but so does the ensemble.

We ran other cases in which just a single skillful model was added to the ensemble of skill-less models. In Case 5-L1, we used the ensemble of models used for Case L1, retaining only the 8 models with correlation coefficients < 0.3 *and* added Model no. 3, which has a correlation coefficient of 0.91. In Case 5-L2 we took the ensemble of models used for Case L2 with correlation coefficients < 0.5 *and* added Model no. 3, which has correlation coefficient of 0.93 (see table III). The improvement in the results is quite dramatic as shown by fig. 5, which compares the time series with and without the skillful model.

Last, we ran additional cases in which only models of moderate skills with training correlation coefficients between 0.5 and 0.75 were retained in the ensemble (Case 6-L1 and Case 6-L2, using the ensemble of models for Case L1 and L2, respectively). These results are also shown in fig. 5, and can be seen that the performance of SE technique is now intermediate to the above two cases as could be expected.

In all the above cases, the SE technique is consistently more skillful than the simple ensemble average. FSI_{COR} is always greater than zero, significantly in some cases, marginally in others (see table III for the statistics). FSI_{RMS} and FSI_{BIAS} values depend on whether *tmean* or *emean* is close to the true forecast mean.

Thus it is fair to say that no matter how many models are in the ensemble, if they are of questionable skill, the SE technique cannot be expected to yield good forecasts,

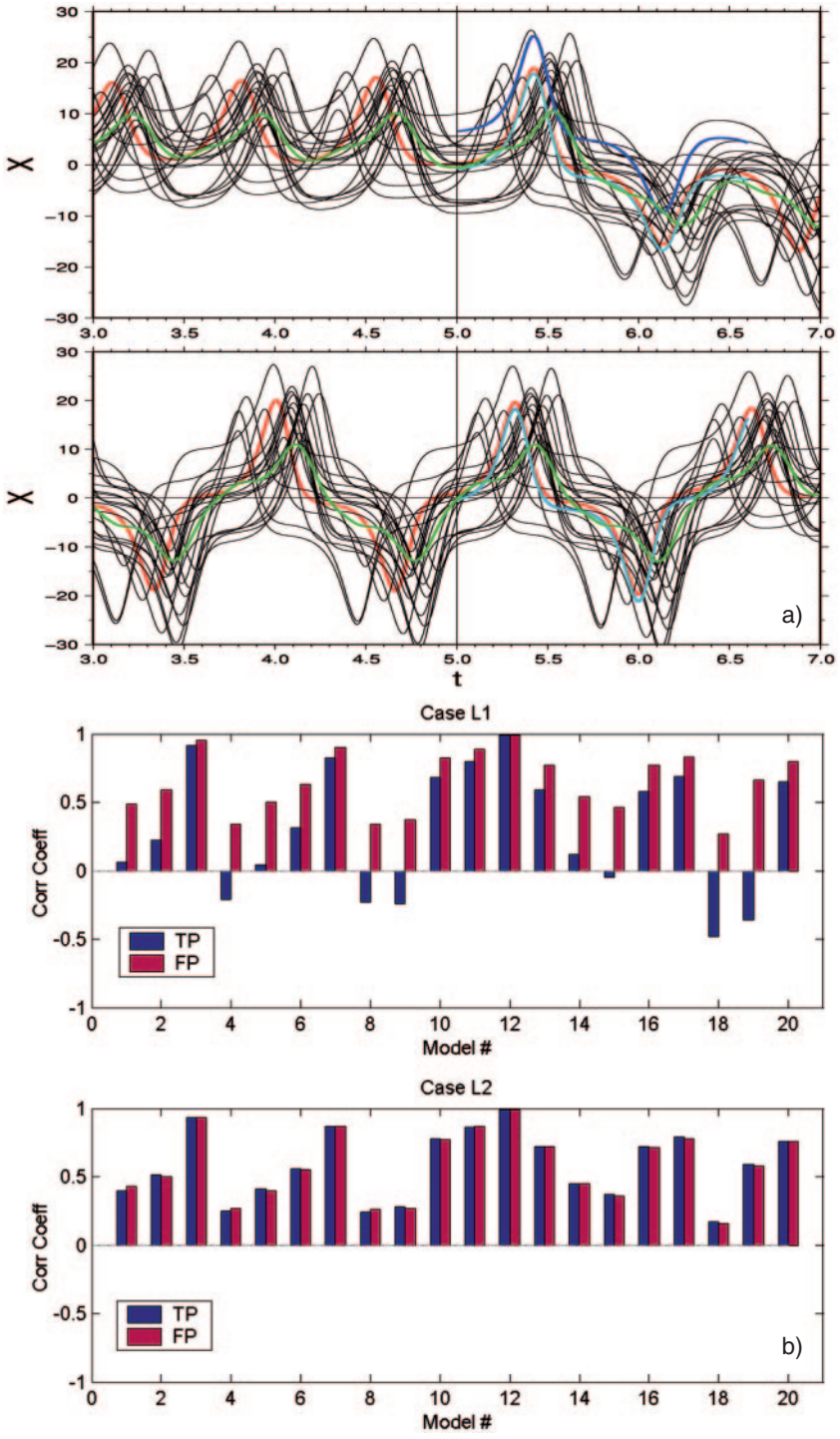

Fig. 3. – a) Time series of individual models (black lines), simple ensemble (green) and superensemble averages (dark blue: *tmean*, light blue: *emean*) for Case L1 (top panel) and Case L2 (bottom panel). Red denotes the "truth", the vertical line the end of the TP. b) The models in the ensemble and their correlation coefficients for Case L1 and L2. TP: training period; FP: forecast period.

TABLE III. – *Statistics for the Lorenz systems. The numbers in the 2nd column report the number and the correlation coefficient of the models adopted in that specific run. The number in parentheses in the rest correspond to the use of* emean *instead of* tmean. *(Only the numbers differing from those of the* tmean *are shown). Observed forecast mean (variance): Case L1:* −0.48 *(*59.41*); Case L2:* 1.28 *(*78.67*).*

Case	N	$VAR_{\rm SE}$	$VAR_{\rm E}$	$M_{\rm SE}$	$M_{\rm E}$	$COR_{\rm SE}$	$COR_{\rm E}$	$RMS_{\rm SE}$	$RMS_{\rm E}$	$BIAS_{\rm SE}$	$BIAS_{\rm E}$	$FSI_{\rm COR}$	$FSI_{\rm RMS}$	$FSI_{\rm BIAS}$
L1	20 (*all*)	59.48	37.26	5.81 (−1.58)	−1.58	1.00	0.79	6.30 (1.10)	4.82	6.30 (−1.10)	−1.10	0.26	−0.24 (3.38)	−0.83 (0.0)
2-L1	12 (> 0.3)	59.03	45.41	5.81 (−0.64)	−0.64	1.00	0.85	6.30 (0.16)	4.01	6.30 (−0.15)	−0.15	0.17	−0.36 (24.2)	−0.98 (0.0)
3-L1	8 (< 0.3)	30.33	34.18	5.81 (−3.00)	−3.00	0.86	0.60	7.50 (4.79)	6.78	6.30 (−252)	−2.52	0.44	−0.96 (0.42)	−0.60 (0.0)
4-L1	2 (> 0.9)	61.40	51.84	5.81 (0.26)	0.26	1.00	0.99	6.30 (0.80)	1.32	6.30 (0.75)	0.75	0.01	−0.79 (0.65)	0.88 (0.0)
5-L1	9 (< 0.3+#3)	52.43	34.01	5.81 (−3.63)	−3.63	0.99	0.68	6.37 (3.30)	6.52	6.30 (−3.15)	−3.15	0.47	0.23 (0.98)	−0.50 (0.0)
6-L1	5 (0.5–0.75)	52.08	49.95	5.81 (2.01)	2.01	0.99	0.81	6.44 (2.84)	5.32	6.30 (2.49)	2.49	0.23	−0.17 (0.87)	−0.60 (0.0)
L2	20 (*all*)	78.69	40.83	−0.14 (−0.27)	−0.27	1.00	0.75	1.42 (1.54)	6.00	−1.42 (−1.54)	−1.54	0.32	3.24 (2.89)	0.09 (0.0)
2-L2	16 (> 0.3)	78.68	49.01	−0.14 (−0.35)	−0.35	1.00	0.78	1.42 (1.63)	5.77	−1.42 (−1.63)	−1.63	0.28	3.07 (2.54)	0.15 (0.0)
3-L2	4 (< 0.3)	28.14	36.01	−0.14 (0.06)	0.06	0.60	0.39	7.26 (7.22)	8.67	−1.42 (−1.21)	−1.21	0.55	0.19 (0.20)	−0.14 (0.0)
4-L2	2 (> 0.9)	79.05	66.08	−0.14 (1.79)	1.79	1.00	0.99	1.46 (0.62)	1.49	−1.42 (0.52)	0.52	0.01	0.25 (1.39)	−0.63 (0.0)
5-L2	8 (< 0.5+#3)	75.81	32.85	−0.14 (−1.55)	−1.55	0.99	0.62	2.06 (3.16)	7.51	−1.42 (−2.82)	−2.82	0.59	2.76 (1.38)	1.00 (0.0)
6-L2	4 (0.5–0.75)	63.63	69.76	−0.14 (−2.43)	−2.43	0.93	0.64	3.58 (4.96)	8.24	−1.42 (−3.71)	−3.71	0.46	1.30 (0.66)	1.62 (0.0)

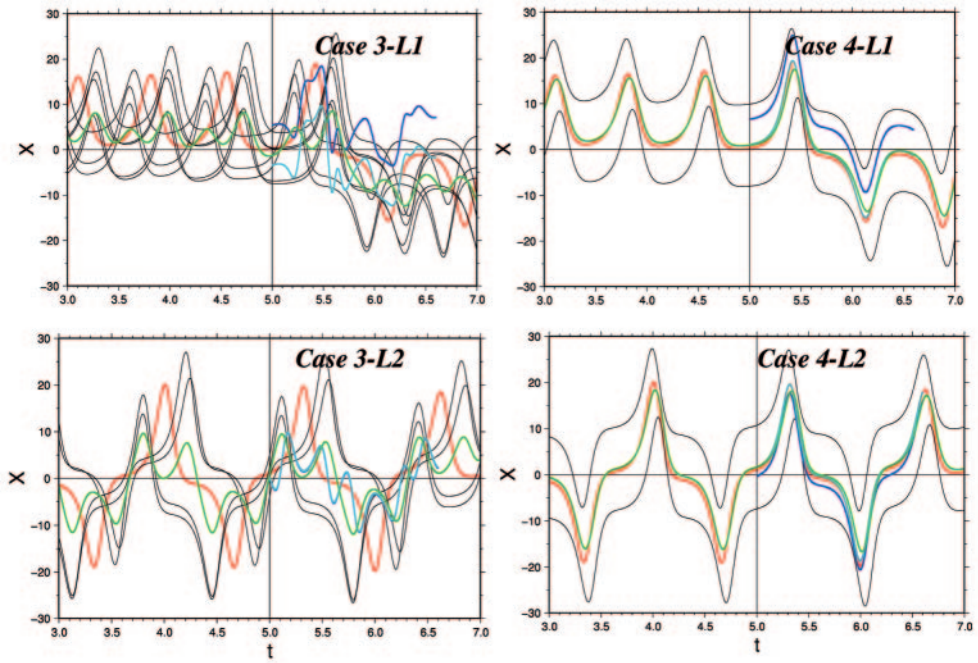

Fig. 4. – Time series of individual models (black lines), simple ensemble (green) and superensemble (dark blue: *tmean*, light blue: *emean*) averages for different cases of Lorenz regimes. Red denotes the "truth", the vertical line the end of the TP (see table III for the corresponding statistics).

even though it will be more skillful overall than the simple ensemble average. On the other hand, when the ensemble includes even a single skillful model, the SE technique provides good forecasts. However, it is not clear *a priori* which mean should be used in the SE forecast: depending on the particular situation, either the *tmean* or the *emean* can provide better results. In any case, an overwhelming outcome of these trials with the Lorenz chaotic system is that it is important that the ensemble include skillful models for the SE technique to be of any practical utility. Naturally, the technique is only as good as the best models in the ensemble.

4. – Concluding remarks

The major strength of the SE technique is its ability to use observations in the past to assess the performance of the individual models in the ensemble, and determine the optimum weighting to be assigned to the individual models to provide the best forecast possible. Consequently, the SE forecast is almost always better than the simple ensemble forecast, although the degree of improvement depends on the properties of the models in the ensemble. However, the skill of the SE forecast with respect to the true forecast depends on a number of factors, principally on the skill of the models in the ensemble. As can be expected, if the ensemble consists of models with poor skill, the SE forecast will also be poor, although better than the simple model ensemble forecast. Inclusion of even a single skillful model in the ensemble can increase the forecast skill significantly.

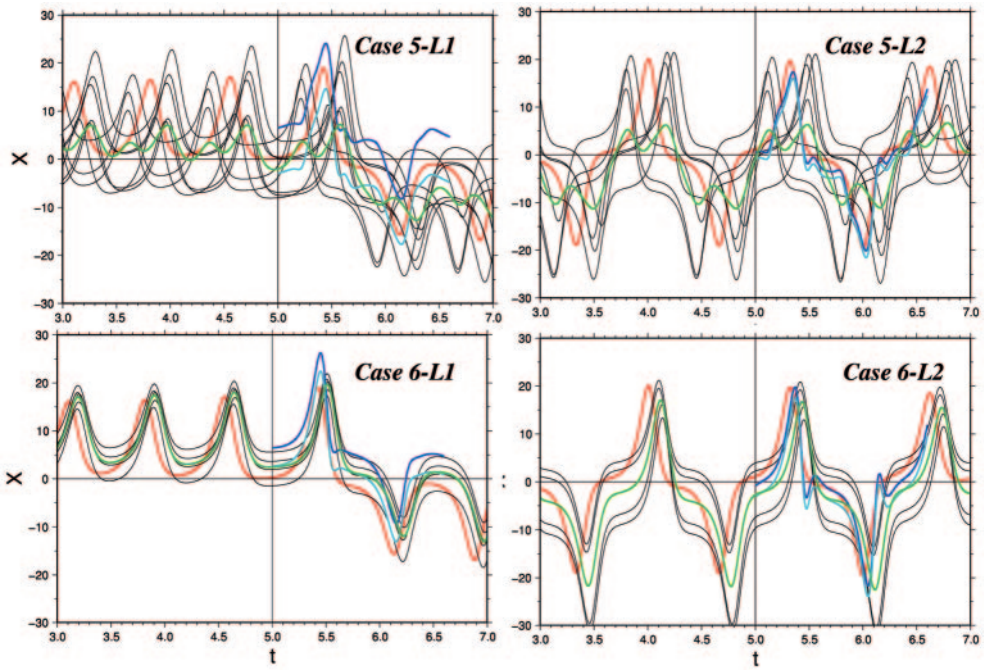

Fig. 5. – Time series of individual models (black lines), simple ensemble (green) and superensemble (dark blue: *tmean*, light blue: *emean*) averages for different cases of Lorenz regimes. Red denotes the "truth", the vertical line the end of the TP (see table III for the corresponding statistics).

The technique does provide a means of assessing the skill of each individual model in the ensemble of forecast models, and this makes it possible to exclude consistently unskillful models from the ensemble, also contributing to saving computing resources.

The obvious weakness of the SE forecast is that the method assumes that error statistics from the past can be used for arriving at the best forecast of the future. Indeed, prior behavior is not necessarily a guide to future performance and this of course could limit the forecast skill in some cases. Nevertheless, the technique is of clear utility in short-term ocean forecasts and operational atmospheric/oceanic forecast centers, both civilian and naval, could benefit significantly from a routine day-to-day use of this technique.

* * *

LK and SC thank ONR and Drs. M. Fiadeiro and S. Harper for partial support for this work through ONR grants N00014-06-10287 and N00014-05-1-0730. The support from the CNR-RSTL "MOM" was highly appreciated.

REFERENCES

[1] Krishnamurti T. N., Kishtawal C. M., LaRow T., Bachiochi D., Zhang Z., Williford E., Gadgil S. and Surendran S., *Science*, **285** (1999) 1548.
[2] Krishnamurti T. N., Kishtawal C. M., Zhang Z., LaRow T., Bachiochi D., Williford E., Gadgil S. and Surendran S., *J. Climate*, **13** (2000) 4196.

[3] Kantha L. H. and Clayson C. A., *Numerical Models of Oceans and Oceanic Processes* (Academic Press) 2000.
[4] Onken R., Robinson A. R., Kantha L. H., Lozano C. J., Haley J. P. and Carniel S., *J. Marine Syst.*, **56** (2005) 45. DOI: 10.1016/j.marsys.2004.09.010.
[5] Yun W. T., Stefanova L. and Krishnamurti T. N., *J. Clim.*, **16** (2003) 3834.
[6] Williford C. E., Krishnamurti T. N., Torres R. C., Cocke S., Christidis Z. and Kumar T. S. V., *Mon. Weather Rev.*, **131** (2003) 1878.
[7] Palmer T. N., Alessandri A., Andersen U., Cantelaube P., Davey M., Delecluse P., Deque M., Diaz E., Doblas-Reyes F. J., Feddersen H., Graham R., Gualdi S., Guérémy J.-F., Hagedorn R., Hoshen M., Keenlyside N., Latif M., Lazar A., Maisonnave E., Marletto V., Morse A. P., Orfila B., Rogel P., Terres J.-M. and Thomson M. C., *Bull. Am. Meteorol. Soc.*, **85** (2004) 853.
[8] Rixen M. and Ferreira-Coelho E., *Ocean Modelling*, **11** (2006) 428.
[9] Press W. H., Teukolsky S. A., Vettering W. T. and Flannery B. P., *Numerical Recipes in Fortran*, 2nd edition (Cambridge University Press) 1992.
[10] Lorenz E. N., *J. Atmos. Sci.*, **20** (1963) 130.