

RELATIONSHIP BETWEEN ERROR AND ENSEMBLE SPREAD IN A REGIONAL ENSEMBLE FORECAST SYSTEM FOR SOUTH AMERICA

Juan J. Ruiz^{1,2}, Celeste Saulo^{1,2}, Eugenia Kalnay³

¹University of Buenos Aires, Buenos Aires, Argentina ²Centro de Investigaciones del Mar y de la Atmósfera (CONICET-UBA), Buenos Aires, Argentina ³University of Maryland, College Park, USA.

1. INTRODUCTION

According to Kalnay (2003) ensemble forecasting has two major advantages, one is to provide a better estimation of the forecasted variables through the use of the ensemble mean, and the other is the estimation of the forecast uncertainty through the ensemble spread. Though originally developed for medium range global forecasts, short range regional forecasts can also benefit from ensemble forecasting techniques (see for example Hou et al. 2001). Little is known about the performance of regional ensemble forecasts over South America in terms of error reduction and forecast uncertainty estimation (Silva Dias et al. 2006, and Ruiz et al. 2006). This work provides an assessment of the performance of a regional and a global ensemble system during the 2002-2003 warm season.

Several works studied the relationship between error and ensemble spread: Houtekamer (1993) develops a theoretical model for this relationship, and Whitaker and Lounge (1998) analyze this relationship for a global ensemble over the Northern Hemisphere. In these works a linear correlation coefficient was used to measure the strength of the relationship between the two variables. Grimit and Mass (2007), use the theoretical model provided by Houtekamer (1993) to test other measures of strength of this relationship, most of them not based on a linear relationship between the two variables. In the present work we propose to study some aspects of the relationship between error and spread without restricting ourselves to the a linear framework.

2. METODOLOGY

2.1 GLOBAL AND REGIONAL ENSEMBLES

Two short range ensembles - one global and the other one regional- are studied and compared in this work. The global ensemble uses the MRF (Medium Range Forecasts) model with T62L28 resolution (approximately 2.5° horizontal resolution). The breeding of the growing modes technique (hereafter breeding) (Toth and Kalnay 1993) is used to introduce perturbations in the initial conditions with a rescaling period of 6 hours. The ensemble consists of 11 members (5 pairs of perturbed members and a control run) integrated up to 48 hours lead time.

This global ensemble is also used to provide initial and boundary conditions to a regional ensemble based on the WRF model version 2.0 (Skamarock et. al. 2005) which has been run with 40 km horizontal resolution and 31 sigma vertical levels. The convective parameterization selected is Kain-Fritsch (Kain 2004), the boundary layer parameterization is the Yonsei University scheme (Hong and Pan 1996), and the surface processes are modeled using the NOAH surface model (Dudhia 2001). The regional ensemble has the same number of members as the global ensemble, and each member of the regional model is nested in its corresponding global ensemble member and is integrated up to 48 hours lead time.

Both ensembles were initialized twice a day at 00 and 12 UTC: the global ensemble uses NCEP-NCAR Reanalysis (Kalnay et al. 1996) as unperturbed initial condition while the regional ensemble uses the Global Data Assimilation System analysis with a resolution of 1°x1°. The experiment starts on 15th December 2002 and ends on 15th February 2003.

2.2 ERROR COMPUTATION

To compare the magnitude of the ensemble mean error with that of the control forecast for both ensembles, the error time average (bias) was first computed at each grid point (Equation 1), where f_i is the forecasted value at each grid point for the i^{th} time and o_i is its corresponding observation. This is done in order to remove part of the systematic error from

Corresponding author address: Juan J. Ruiz, CIMA/University of Buenos Aires, Ciudad Universitaria, Buenos Aires, Argentina.
E-mail: jruiz@cima.fcen.uba.ar

both models before performing a comparison among them. The mean absolute error for each grid point is then computed as in Equation 2.

$$bias = [f_i - o_i] \quad (1)$$

$$MAE = \frac{1}{N} \sum_{i=1}^N |f_i - o_i - bias| \quad (2)$$

The Japanese Meteorological Agency Reanalysis (JRA) (Onogi et al. 2007) with a horizontal resolution of $2.5 \times 2.5^\circ$ are used to compute the forecast error. This dataset is independent of the ones used to initialize both ensembles. Also data from surface weather stations available through the Global Telecommunication System were used for error computation; these data have been interpolated to the global model grid through box averaging. To compare the error between ensemble systems the WRF output was also interpolated to the MRF grid using the same technique (since the MRF resolution was much lower). In this way both models are compared at horizontal scales that can be resolved by both of them (although the MRF model might have a horizontal effective resolution which is lower than 2.5°).

2.3 METRICS TO EVALUATE THE RELATIONSHIP BETWEEN THE ENSEMBLE MEAN AND SPREAD

According to Grit and Mass (2007) the relationship between the ensemble error and spread can be studied without assuming a linear relationship between them. Instead of that, the joint distribution of those variables is examined using scores based on the Continuous Ranked Probability Score (Wilks 1995). In this work both, the spatial and temporal dependence of the relationship, are examined.

As the considered region goes from the mid latitudes to the tropics and from the eastern Pacific to the wet Amazonian rain forest the values of error and spread should be standardized to allow a comparison between those regions. To do this, the mean and the standard deviation of the absolute error and the spread are computed at each grid point, and those values are used in the standardization process. The mean and the standard deviation are computed independently for each initialization time and for each forecast lead time, this is done in order to take into account that forecasts with different initialization time verify at

different times of the day and as the error and spread have a strong diurnal cycle (mainly below 850hPa) the error-spread relationship would be dominated by the strong diurnal cycle.

Since the idea is to analyze the spatial distribution of the relationship strength between error and spread, one of the most important limitations is the sample size. As the experiment is approximately 60 days long, we have 120 error-spread pairs at each grid point and for each forecast lead time which is not enough to study the joint distribution of both variables. To increase the sample size a procedure similar to that proposed by Cussack and Arribas 2008 is followed. The sample at each grid point is composed by the error-spread pairs corresponding to that particular grid point plus the pairs corresponding to the 8 surrounding grid points. In this way, the sample size at each grid point was increased to 1080. Using this approach, two adjacent grid points share 66% of the sample so the results become more dependent. Furthermore, in regions where the mean spatial gradient of error and spread are strong this could lead to a spurious relationship between them. In this work, this particular issue has been addressed by the standardization of both variables with respect to their local mean and standard deviation.

To quantify the strength of the relationship between error and spread the spread is first discretized into 4 categories, determined by the quartiles of the probability distribution of both variables at each grid point. Also 3 thresholds are defined for the error based on the quartiles of the error distribution at each grid point (i.e. errors above the lowest threshold have a probability of occurrence of 75% and errors above the highest threshold have a probability of occurrence of 25%). The probabilities for errors above each threshold can also be computed independently for each spread category. If a relationship between error and spread exists, then the probability of having an error above a certain threshold will increase if categories corresponding to higher spread values are considered. In Figure 1, the red line represents the a priori probability (P_c) of having an error above a value corresponding to the 50% probability. This value is independent of the spread category considered. The red dashed line represents the conditional probability of having an error above the 50% threshold as a function of the spread category considered. The blue line represents an ideal situation: in this case when the spread is in the first or the second quartile, then the probability of having an error above a value corresponding to the 50% probability is null, and is one if the spread is on the third or fourth quartile. Based on this figure and on the idea of the

continuous ranked probability score a Resolution Index is defined to measure the strength of the relationship between error and spread (Equation 3), where P_d is the probability corresponding to the perfect deterministic relationship between the two variables, (the blue line in Figure 1) and P_{cs} is the observed frequency of occurrence at each spread category k . The resolution index can also be computed for the hypothetical situation where no relationship exists between the two variables as in Equation 4, where P_c is the a priori probability of occurrence of errors above the considered threshold. Combining these two definitions a Resolution skill score (RSS) can be defined as in Equation 5, the maximum (better) value of the score is one, negative values indicate values of R greater than R_c : this could sometimes mean that the relationship between the considered variables is not direct (which is not the case in this work).

$$R = \sum_k (p_{cs}(k) - p_d(k))^2 \quad (3)$$

$$R_c = \sum_k (p_c - p_d(k))^2 \quad (4)$$

$$RSS = 1 - \frac{R}{R_c} \quad (5)$$

There is no parametric test to evaluate the statistical significance of the RSS for each grid point under the assumptions of this work, so to evaluate the significance, 10^4 random samples were generated assuming a persistence with a temporal correlation coefficient of 0.95 at lag 1 and a spatial correlation coefficient of 0.95. This should take into account the fact that the error series from neighbor grid points are not independent. The RSS was computed for each series (each with the same size as the series available from the experiment) and 99% of the series were below 0.26 and 95% of them were below 0.21. Consistently, values higher than those indicate the existence of a relationship between error and spread which is statistically significant. In the experiments considered in this work, the spatial correlation coefficient of the errors depends on the variable under consideration: for some of them, the spatial correlation coefficient is lower than the one used for the random series generation so the significance thresholds computed in this section should be taken as an upper limit.

3. RESULTS

3.1 ERROR OF THE CONTROL RUN VS THE ENSEMBLE MEAN ERROR.

The ensemble mean error for both ensembles was compared with the control forecast error over South America. Figure 2, shows an example of the averaged error reduction (%) of the ensemble mean with respect to the error in the control forecast. In this example – corresponding to 2 meter specific humidity (q_2m) - it can be seen that ensemble mean errors are 10-20% lower than the control run errors, and the error reduction is more important at 48 hours lead time. A similar Figure was constructed using data from the available surface stations and the results were quite similar both in the spatial distribution and in the error magnitude (not shown). The most important errors in the moisture field are located over Northern Argentina where q_2m variability is also maximum due to the presence of a strong humidity gradient between the tropics and the extra-tropics.

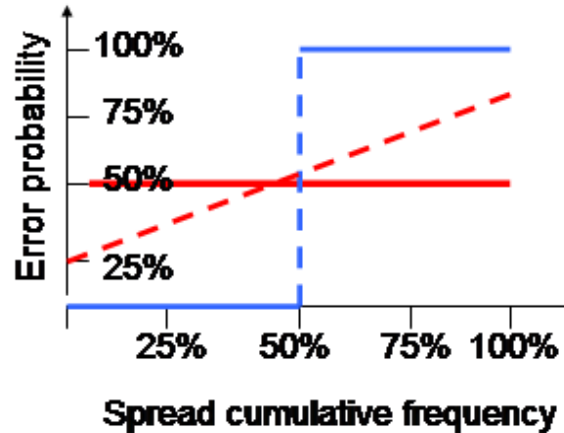


Figure 1: Schematic illustration of possible behaviours of the error-spread joint distribution. See Figure description in the text.

To summarize the behavior of q_2m , T_2m and SLP , Figure 3, shows the evolution of the error and the spread for these 3 variables as a function of the forecast lead time and for 3 selected regions indicated by the red squares in Figure 2. Region 1 was selected as an example of a region with high model error: at this particular place both models have problems related to the underestimation of the strength of the low level jet. The second region, located to the north is an example of a tropical regime, where forecasts errors are small, usually because errors saturate at very short forecast lead

time and at small values. The third region was selected because it is located at mid latitudes

and is affected by baroclinic waves.

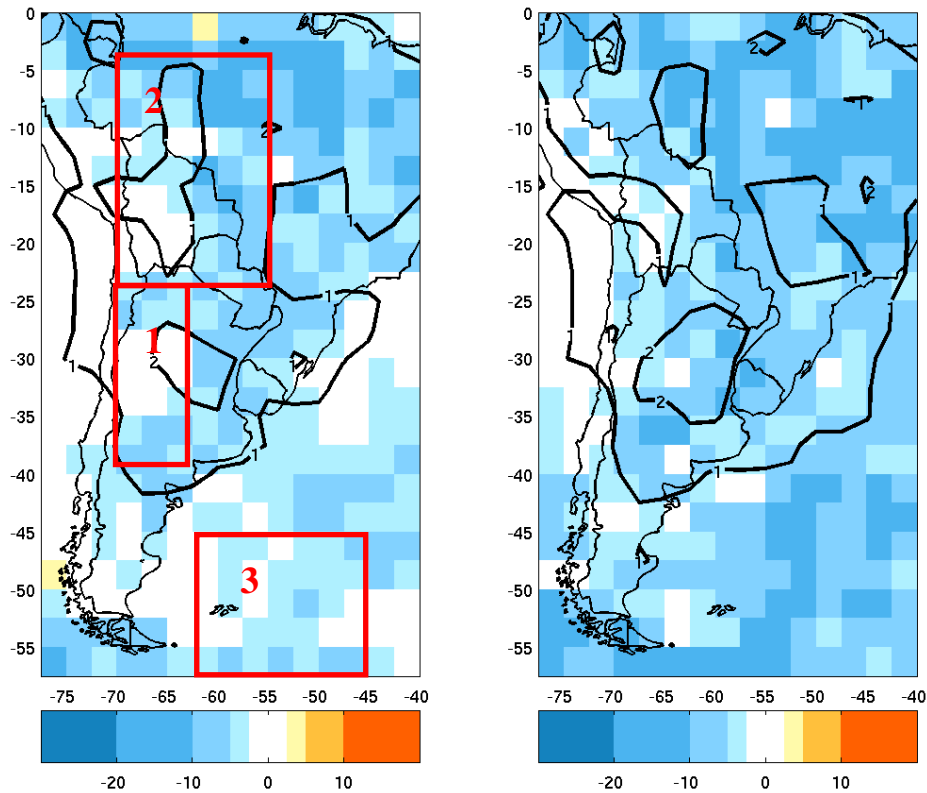


Figure 2: q2m error difference between the ensemble mean and the control forecast (shaded) (in % with respect to the control forecast error) and control forecast error (g Kg⁻¹) (contours) for 24-hr forecast (left) and 48-hr forecast (right). Numbered boxes show the position of the regions that are analyzed in the text.

As can be seen in Figure 3 the error for the ensemble mean is smaller for the regional ensemble than for the global ensemble almost everywhere. This can be due either to the higher resolution of the regional ensemble or to the increased resolution of its initial conditions. However, the ensemble spread is also smaller while the spread growth rate is almost the same in both cases. The behavior of the error and the spread inside the 3 regions is different: in region 1, where the model error is larger, the difference between the error in the mean and the ensemble spread is maximum. The ensemble is under dispersive almost everywhere, but particularly at this region. Region 2 shows almost null error and spread growth rate as expected due to the forecast error behavior in the tropics (Kalnay 2003). Region 3 shows the fastest error and spread growth for the SLP field mostly due to

baroclinic activity in this area. On the other hand, the error growth for the moisture field is smaller than in the other two regions: this is because the moisture content is smaller over Region 3. Region 3 is where spread is closer to forecast error, probably because both, spread and forecast error, are dominated by perturbations that grow because of baroclinic instability.

These Figures show that both ensembles can reduce forecast error and that the ensemble spread behaves in a similar way as the forecast error when their increase with forecast lead time is analyzed. However this does not guarantee that a day to day relation between these variables exists. The question that will be answered next is if larger dispersion is directly associated with larger forecast error.

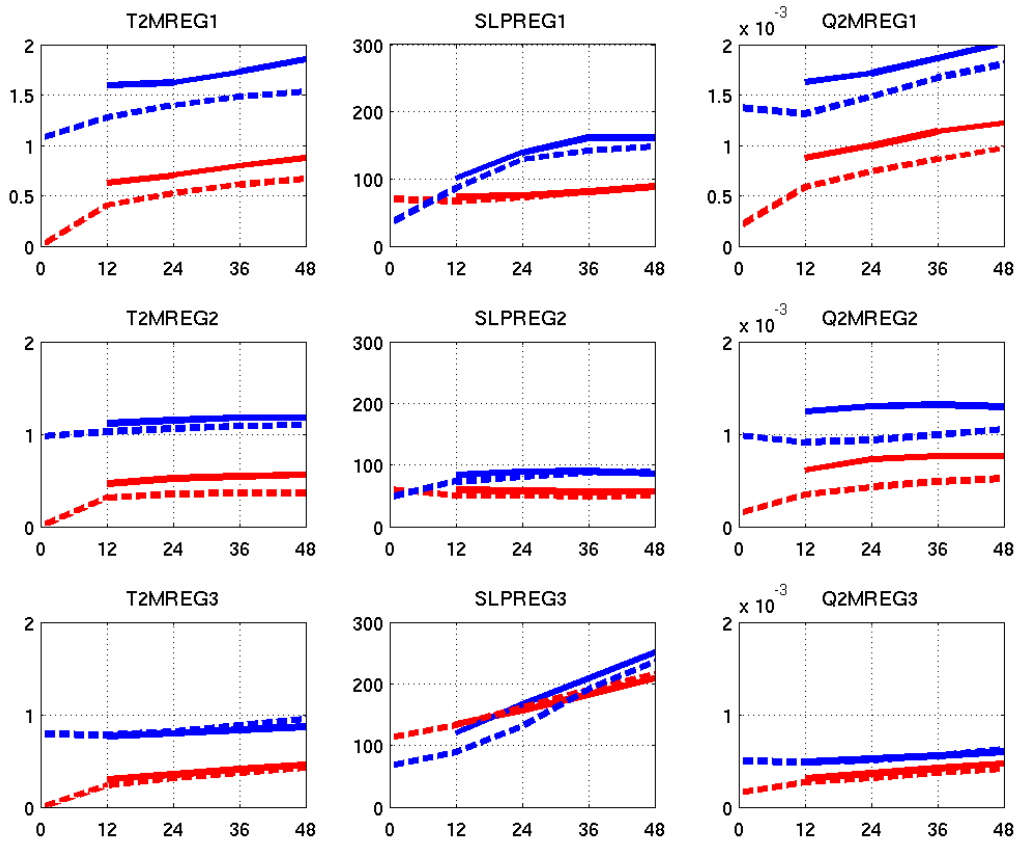


Figure 3: Averaged error (blue lines) and dispersion (red lines) for the global ensemble (solid lines) and the regional ensemble (dashed lines) over the different regions described in the text (region 1, first row, region 2 second row and region 3 third row), for 2-m temperature (first column), sea level pressure (second column) and 2-m moisture (third column).

3.2 RELATIONSHIP BETWEEN ERROR AND SPREAD

To explore the day to day relationship between the forecast error and the ensemble spread the standardized spread range is divided into 4 categories (i.e. the quartiles of the standardized spread distribution) as described in section 2. At each category the probability of occurrence of different error thresholds is also computed. These thresholds were computed using the quartiles of the standardized error distribution. Figure 4, shows the results for the sea level pressure at the 3 regions previously described. As can be seen from the figure, there is no relationship between the model error and the ensemble spread for SLP at Region 2. This is also true for other variables in this region. Over region 1 there is some relationship between spread and error at 48 hour lead time. The only region that shows a clear relationship between

error and spread for all forecast length and the strength of this relationship seems to increase with forecast lead time is region 3. The slope of the curves in this region also suggests that the relationship between error and spread is stronger in the presence of large errors in close agreement with the results of Houtekamer (1993) and Grit and Mass (2007).

To study more in detail the spatial distribution of the strength of the relationship between spread and error over South America the RSS has been computed at each grid point using data from the adjacent grid points as described in section 2. Figure 5 shows the results obtained for the sea level pressure for three different error thresholds (25%, 50% and 75%) and for 3 lead times. As expected for this variable, the higher RSS values take place at mid-latitudes, where baroclinic activity is stronger. The strength of the relationship is maximum over the Atlantic, where also the spread temporal variability reaches its maximum, which is an expected result

according to Houtekamer (1993). As has been described in previous figures, the strength of the relationship increases with forecast lead time from weak values at the beginning of the

forecast period. As described in Houtekamer (1993) there is also a stronger relationship between both variables when higher error threshold is considered.

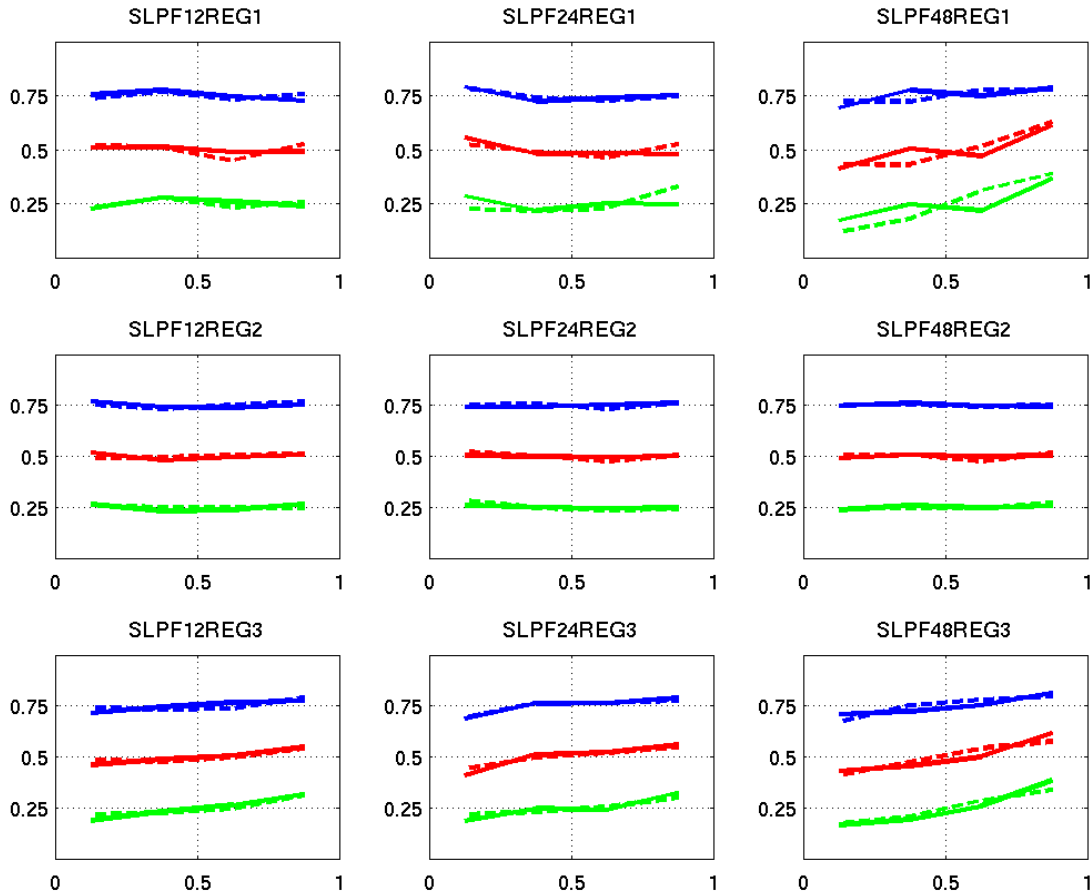


Figure 4: Error probability for the sea level pressure (green (/red/blue) lines: errors with 25% (50%/75%) probability of occurrence) as a function of ensemble spread cumulative frequency. Solid lines for the global ensemble and dashed lines for the regional ensemble. Region 1, first row; region 2, second row and region 3, third row. Column 1, 12 hs lead time, column 2, 24 hours lead time and column 3, 48 hours lead time.

Figure 6, shows the results for moisture at the 950 hPa. level. In this case the spatial distribution is similar, but the strength of the relationship is larger particularly at mid-latitudes. This might be due to the fact that at mid and high latitudes a local maximum in the moisture content is associated with larger spread and forecast errors. This effect is particularly important at higher levels where moisture errors seems to be strongly associated with the

ensemble spread because error and spread are larger in areas where the moisture content is relative large.

The global ensemble shows for both variables similar spatial patterns for the strength of the relationship between spread and error, however the strength of this relationship in the regional ensemble seems to be slightly weaker than in the global case (not shown).

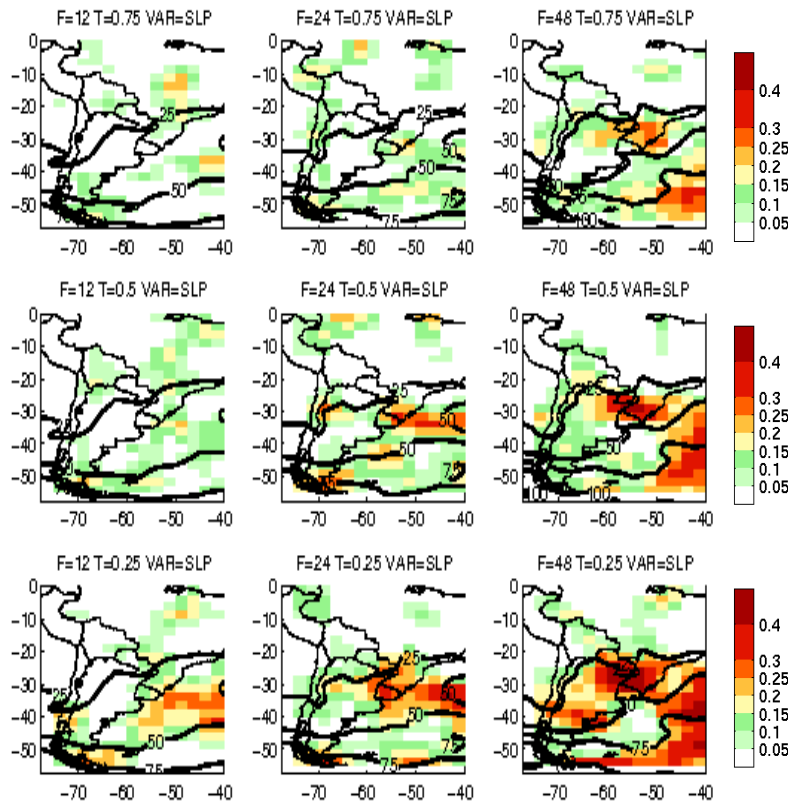


Figure 5: RSS (shaded) for sea level pressure (regional ensemble) and spread variability (contours in Pa). First row, errors with a climatological probability of 75%, second row, 50%, third row 25%). First column, 12 hours lead time, second column 24 hours lead time and third column 48 hours lead time.

CONCLUSIONS

In this work a global and a regional ensemble based on the breeding technique have been verified and compared over South America. Both ensembles show an error reduction in the order of 10-20% with respect to the control run; however, the regional ensemble shows lower errors.

The relationship between error and spread for both ensembles has been also examined with encouraging results. Yet, the regional ensemble seems to show a weaker relationship between error and spread. This problem can be reduced by applying breeding within the regional domain instead of using the perturbations generated by the global model, potentially leading to the generation of perturbations that could grow faster in the regional domain and that could be more consistent with the mesoscale model dynamics.

The use of a non linear approach to describe the relationship between both variables works well in this context and can be applied to the more general problem of model verification allowing for a description of the behavior of different variables under diverse conditions (i.e. the extreme values can be examined using the proper threshold for the variable).

ACKNOWLEDGMENTS

The authors are thankful to Istvan Szunyogh for providing the scripts to run the global ensemble, to Erick Kostelich for his help with the MRF model runs and to Jae Schemm for providing the initial conditions in the required file format.

This study has been supported by the following projects: ANPCyT PICT 2004 25269, UBACyT X204, CONICET PIP 5417 and GC06-085 from NOAA/OGP/CPA.

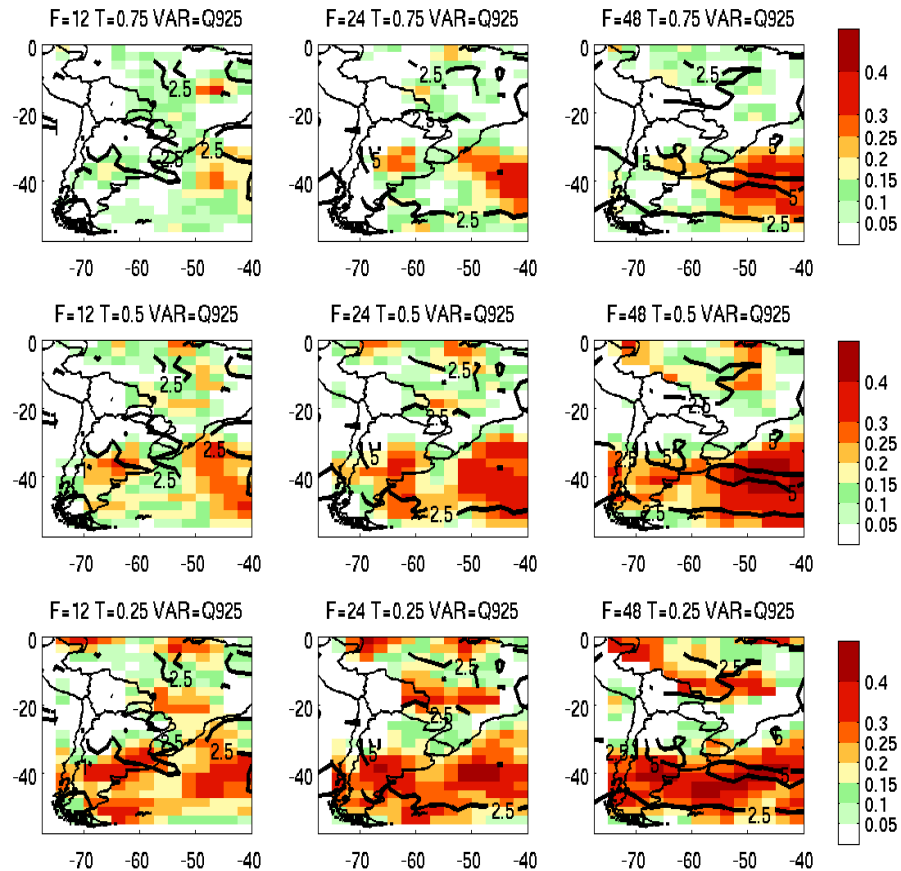


Figure 6: As in Figure 5, but for the specific humidity at 925 hPa.

REFERENCES

Cusack, S. and A. Arribas, 2008: Assessing the Usefulness of Probabilistic Forecasts. *Mon. Wea. Rev.*, **136**, 1492-1504.

Grimit and C. Mass, 2007: Measuring the Ensemble Spread-Error Relationship with a Probabilistic Approach: Stochastic Ensemble Results. *Mon. Wea. Rev.*, **135**, 203-221.

Hong S. and H. Pan, 1996: Nonlocal Boundary Layer Vertical Diffusion in a Medium-Range Forecast Model. *Mon. Wea. Rev.*, **10**, 2322-2339.

Hou, D., E. Kalnay, and K. K. Droegemeier, 2001: Objective Verification of the SAMEX'98 Ensemble Forecast. *Mon. Wea. Rev.*, **129**, 73-91.

Houtekamer, P., 1993: Global and Local Skill Forecasts. *Mon. Wea. Rev.*, **121**, 1834-1846.

Kain J. S., 2004: The Kain-Fritsch Convective Parameterization: An Update. *J. Appl. Meteor.* **43**, 170-181.

Kalnay, E., 2003: Atmospheric Modeling, Data Assimilation and Predictability. Cambridge University Press.

Kalnay E., M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, J. Woollen, Y. Zhu, A. Leetmaa, R. Reynolds, M. Chelliah, W. Ebisuzaki, W. Higgins, J. Janowiak, K. C. Mo, C. Roplewski, J. Wang, Roy Jenne, and Dennis Joseph, 1996: The NCEP/NCAR 40-Year Reanalysis Project. *Bull. Amer. Meteor. Soc.*, **77**, 437-471.

Onogi, K., J. Tsutsui, H. Koide, M. Sakamoto, S. Kobayashi, H. Hatsushika, T. Matsumoto, N. Yamazaki, H. Kamahori, K. Takahashi, S. Kadokura, K. Wada, K. Kato, R. Oyama, T. Ose, N. Mannoji and R. Taira, 2007: The JRA-25 Reanalysis. *J. Meteor. Soc. Japan*, **85**, 369-432.

Ruiz J. J., A. C. Saulo and E. Kalnay, 2006: A regional ensemble forecast system for southeastern south America: preliminary assessment. Proceedings of 8 ICSHMO, Foz do Iguaçu, Brazil, April 24-28, INPE, p 1977-1984.

Silva Dias, P. L., D. Soares Moreira, and G. D. Neto, 2006: The MASTER Model Ensemble System (MSMES). Proceedings of 8 ICSHMO, Foz do Iguaçu, Brazil, April 24-28, INPE, p. 1751-1757.

Skamarock, W. C., J. B. Klemp, J. Dudhia, D. O. Gill, D. M. Barker, W. Wang, and J. G. Powers,

2005: A description of the Advanced Research WRF Version 2. NCAR Tech Notes-468+STR

Toth, Z. and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteor. Soc.*, **74**, 2317-2330.

Whitaker J. and A. F. Loughe, 1998: The Relationship between Ensemble Spread and Ensemble Mean Skill. *Mon. Wea. Rev.*, **126**, 3292-3302.

Wilks, D. S., 1995: Statistical Methods in the Atmospheric Sciences: An Introduction. International Geophysics Series, Vol. 59, Academic Press, 467 pp.