Post-Processing Through Linear Regression.

Bert Van Schaeybroeck and Stéphane Vannitsem

Koninklijk Meteorologisch Instituut (KMI), Ringlaan 3, B-1180 Brussels, Belgium (Contact: bertvs@oma.be)

1 Abstract

We present a comparison of various post-processing schemes for ensemble forecasts, all based on linear regression between forecast data and observations. In order for the regression to be useful in practice, we put forward three criteria which are related to forecast errors, the correct climatological variability and multicollinearity. The regression schemes under consideration include the ordinary least squares (OLS) method, a new time-dependent Tikhonov regression (TDTR), the total least squares (TLS) method, a new geometric mean regression (GM), a error-in-variables (EVMOS) method which was recently proposed by Vannitsem (2009), and finally, a "best member" OLS method (Unger et al., 2009). We find that the EVMOS, the TDTR and GM schemes satisfy all three criteria.

We clarify our theoretical findings using the Lorenz 1963 model. For short lead times, the amount and choice of predictors is more important than the regression method. At intermediate timescales linear regression is unable to provide corrections to the forecast. However, at long timescales the different regression schemes differ strongly and, in order to obtain physically relevant results, the use of OLS should be avoided.

2 Regression

Why Regression?

Meteorological forecasts are subject to errors which originate from model errors and initial-condition errors. To estimate the impact of such errors, ensemble forecasts are generated. Ensemble predictions not only provide the forecaster with a forecast (the mean of the ensemble) but also with an estimate of its variability and therefore its reliability. Forecast skill may be improved by use of statistical post-processing using, for example, linear regression (Glahn and Lowry 1972) which recently has also become highly relevant as attempts are made to combine short-term climate forecasts generated using different forecast models. Postprocessing consists of two steps: 1) regression is applied between forecast and observations using past data. 2) the derived regression parameters are used to correct new forecasts.

Linear Regression

Consider:

- N measurements for the variable X (for instance temperature).
- For each measurement we run our model using slightly perturbed initial conditions. Each ensemble member consists of an (uncorrected) forecast variable V_1 and other model variables V_p ($p = 2, .., P$). We call all V_p's predictors.

The regression problem: use linear regression to optimally combine measurement data X and the forecasted data V_p. Find all regression coefficients β_p such that:

$$X \approx \sum_{p=1}^{P} V_p \beta_p. \tag{1}$$

The near equality is achieved by minimization of some cost function $\mathcal{J}$, different for each regression method and a function of the following errors:

- for each measurement the deviation of measurement data from its value according to the regression function is:

$$\varepsilon_X = X - \sum_{p=1}^{P} \xi_p \beta_p. \tag{2}$$

Here ξ_p is a corrected predictor associated with the predictor V_p.

- for each forecast, we define the forecast uncertainty

$$\varepsilon_{V,p} = V_p - \xi_p. \tag{3}$$

What is a "good" regression method?

Ordinary Linear Regression is the classical approach of regression and is based on ordinary least square (OLS) minimization. However, it has some deficiencies when using it in the context of ensemble forecasts.

We assess the usefulness of a regression method by the following three criteria:

1. The method corrects forecast errors.
2. The method can cope with several highly-correlated predictors which may give rise to multicollinearity.
3. The regression data features the correct climatological variability at long lead time.

Criterion 3: Behavior at long lead time

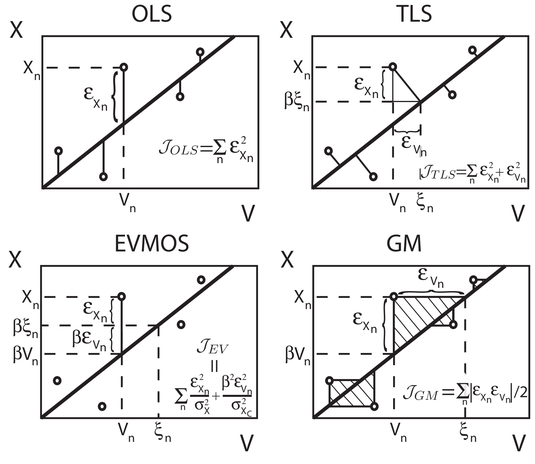
It is well-known that forecasts generated using Ordinary Linear Regression (OLS) converge to the climatological mean at long lead times. However, a forecast with no variability is not physically meaningful. Regression cannot decrease the forecast error at long lead times but could nevertheless yield the "correct" climatological variability. The variability σ_{X_C} of a well-corrected forecast X_C should equal the climatological variability of the measurement data at long lead times:

$$\sigma_{X_C}(t \to \infty) = \sigma_X(t \to \infty). \tag{4}$$

Or, if possible, the regression method may satisfy an even stronger criterion:

$$\sigma_{X_C}(t) = \sigma_X(t), \tag{5}$$

for all times t.



The regression models

The regression schemes and their associated cost functions $\mathcal{J}$ are (see figure above):

1. Ordinary Least Squares (OLS): $\mathcal{J} = \langle \varepsilon_X^2 \rangle$. The bracket denotes the average over all ensemble members.
2. New: Time-Dependent Tikhonov Regression (TDTR): $\mathcal{J} = \left\langle \varepsilon_X^2 + \gamma(t) \sum_p (\beta_p - \beta_p^0)^2 \right\rangle$. Here $\gamma(t)$ is small at small lead times and $\gamma(\infty) \to \infty$.
3. Total Least Squares (TLS): $\mathcal{J} = \left\langle \varepsilon_X^2 + \sum_p (w_p \varepsilon_{V,p})^2 \right\rangle$.
4. Error-in Variables method (EVMOS, Vannitsem 2009): $\mathcal{J} = \left\langle \varepsilon_X^2 + \left(\sum_p \beta_p \varepsilon_{V,p} / \sigma_{X_C} \right)^2 \right\rangle$.
5. New: Geometric Mean (GM): $\mathcal{J} = \left\langle \prod_p |\varepsilon_{V,p} \varepsilon_X| ^{1/P} \right\rangle$.
6. Best-member regression (EREG II, Unger et al. 2009): The same as OLS except the ensemble mean is used and random noise is added.

Except for GM, we have analytical solutions for all methods. In the table below we show the assessment of the different regression methods.

	criterion (1)	criterion (2)	crit.(3), Eq. (4)	strong crit.(3), Eq. (5)
OLS	+	-	-	-
TLS	-	- (+)	-	-
TDTR	+	+	+	-
EVMOS	+	+	+	+
GM	-	+	+	+
EREG II	-	- (+)	+	+

3 Numerical Results

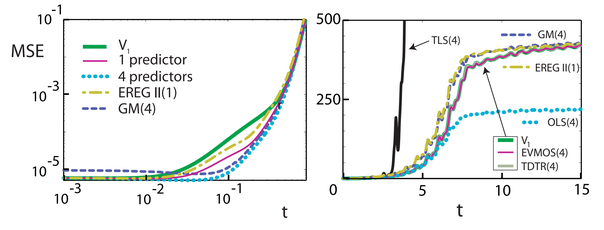
We test the usefulness of the regression methods against the well-known Lorenz 63 model by focussing on the statistical features of the error distributions. The system involves three coupled first-order differential equations in time for the variables x, y and z. We introduce both model and initial-condition errors which have a "comparable" impact on the dynamics. For generating the measurement data we assume a slightly biased parameter set from the one used for generating the model data. We probe the following error variables:

$$u_Z = z - z_C, \quad \text{and} \quad u_r = \sqrt{(x - x_C)^2 + (y - y_C)^2 + (z - z_C)^2}. \tag{6}$$

The Mean Square Error (MSE) Evolution

In the figure below we show the time evolution of the total MSE at short (left) and long (right) lead times (We use 50000 ensembles of each 500 members). At short lead times:

- the corrections are substantial and increase for increasing model errors (Vannitsem and Nicolis 2008).
- the amount and choice of predictors, rather than the regression method itself is of crucial importance.
- GM does not always correct the forecast.



At intermediate lead times:

- there is a fast increase of MSE due to chaotic nature.
- the original forecast V_1 is hardly corrected.
- the error distribution has power-law behavior for large errors.

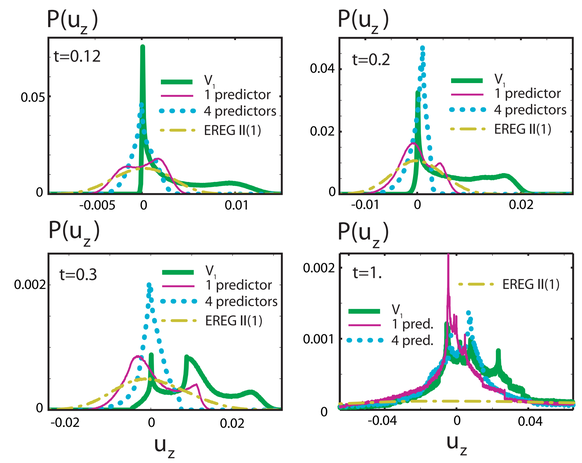
At long lead times:

- the error variance saturates and regression methods strongly differ.
- TLS, GM and EREG II do not always correct the forecast.
- the MSE of the EREG II, EVMOS, TDTR and GM forecasts converge to the value $2\sigma_X^2$ in agreement with criterion (3).
- the MSE of OLS is too low by a factor two.

Evolution of Error Distribution

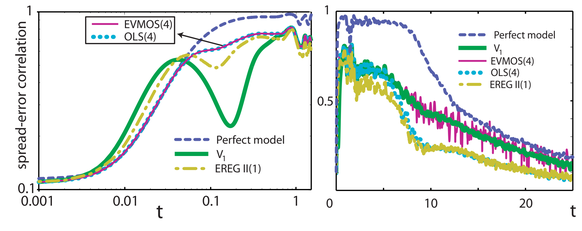
In the figure below we plot the error probability distribution evolution of u_Z (see Eq. (6)) for the Lorenz model *without* initial condition errors. It is clear that:

- the regression quality at short times depends strongly on the number (and choice) of predictors.
- the biased distribution of the uncorrected forecast V_1 (green line) becomes unbiased after post-processing. We show the one-predictor, four-predictor and EREG II distributions.
- four predictors clearly lead to the smallest second moment of the distribution.
- at intermediate times ($t = 1.$) the EREG II distribution becomes very broad compared to the others.



Use of Ensembles

The ensemble spread-error correlation against time is shown in the figures below and is marked by an increased correlation at short times (left figure) for the post-processed forecasts. As expected, a progressive correlation decrease sets in for all ensembles around lead times $t = 5$ (right figure). Remarkably, the OLS and EREG II correlations are distinctly smaller than the ones of the uncorrected and EVMOS ensembles. At $t = 15$, the variance of ensemble spread for all except the EREG II ensembles is still significant.



Prospects

Further details can be found in Van Schaeybroeck and Vannitsem (2011). Current investigations are ongoing applying these post-processing methods on meteorological data from the YOTC project to study the possible usefulness of forecast tendencies as predictors.

Work supported by the Belgian Federal Science Policy Program under contract MO/34/020.

References

- Glahn, H.R., and D.A. Lowry, 1972: The Use of Model Output Statistics (MOS) in Objective Weather Forecasting. *J. Appl. Meteor.*, **11**, 1203-1211.
- Unger, D. A., H. van den Dool, E. O'Lenic, and D. Collins, 2009: Ensemble Regression. *Mon. Wea. Rev.*, **137**: 2365-2379.
- Vannitsem, S., and C. Nicolis, 2008: Dynamical Properties of Model Output Statistics Forecasts. *Mon. Wea. Rev.*, **136**, 405-419.
- Vannitsem, S., 2009: A unified linear Model Output Statistics scheme for both deterministic and ensemble forecasts, *Quart. J. Roy. Meteorol. Soc.*, **135**: 1801.
- Van Schaeybroeck, B., and S. Vannitsem, 2011: Post-processing through linear regression, *Nonlin. Processes Geophys.*, **18**, 147-160.