On Retrieving Parameters of a Linear Regression Model that Accommodates Error Correlation in Well Sampled Data

Rick E Danielson (@gmail.com), Danielson Associates
Office Inc., Halifax, Nova Scotia, Canada

1. Introduction

It is often the case in geophysics that one truth (a so-called genuine truth) is of interest. However, one can distinguish between a *numerical forecast model*, whose ensemble might seek to bracket the best possible estimate of a true evolution, and a *statistical measurement model* (or regression model), which allows for a flexible definition of truth (a so-called partial truth) in the validation of such forecasts. Validation accommodates this flexibility because it implicitly or explicitly acknowledges that data (both model output and observations) have limited support (e.g., a particular range of scales over which they are sensitive). Any comparison between such data is just a comparison of truth on the *intersection* of these supports (e.g., a still smaller range of scales). Other important constraints on truth arise because a measurement model (like a numerical model) is always an approximation (e.g., it assumes linearity). In general, the assumption that errors in collocated estimates of a geophysical variable could be independent of each other carries the difficult implication that only signal (or truth) is correlated while noise (or error) is not. Because measurement models already involve a constrained or partial truth, with errors that may just be consistent with a broader (genuine) truth, a matching (truth-only) constraint on error correlation might be ideal in principle, but is probably quite rare in practice.

2. Model hierarchy

The need to accommodate cross-correlated errors between numerical forecasts and collocated observations is not matched by an existing framework for doing so and a new measurement model is called for. We seek to add to an existing hierarchy of linear regression models, whose foundation (the so-called signal model) is an assumed linear relationship between the error-free component of two variates. The simplest class of model with error (ordinary or reverse linear regression) allows error in just one variate. Allowing errors in both variates is called errors-in-variables regression. Next, we also seek to distinguish between correlated and uncorrelated error in both variates. The equations for collocations of observational (*I*) and gridded (analysis or forecast; *N*) data are

$$\begin{aligned} \text{in situ} \quad I &= t + \lambda_N \epsilon_I + (1 - \lambda_N)\epsilon_I \\ \text{nowcast} \quad N &= \alpha_N + \beta_N t + \lambda_N \epsilon_I + \epsilon_N \end{aligned}$$

A linear regression solution is defined by the additive (α_N) and multiplicative (β_N) calibration of *N* relative to *I*. Correlated error ($\lambda_N\varepsilon_I$) and uncorrelated errors ($[1\text{-}\lambda_N]\varepsilon_I$ and ε_N) are also included. Analytic solutions (e.g., by the method of moments) is generally only possible for ordinary and reverse linear regression. Again in geophysics, however, there is the opportunity to propose numerical solutions to experimental measurement models that sample large observational datasets as well as high resolution analysis and forecast datasets that are orders of magnitude larger still. Measurement models like the

one above that distinguish between correlated and uncorrelated error are also applied in studies of the human condition, but such studies involve comparatively few people. These days, the opportunity to solve measurement models by an experimental sampling of geophysical data is usually far cheaper.

3. Linear regression with symmetric AR-1 errors

For models with many unknowns, it is common practice to use proxy data (so-called instruments) to supplement the information from the collocated samples (*I* and *N*). Danielson et al. (2018) propose to use a few nearly collocated samples as persistence forecasts (*FE*) and revcasts (*RS*) of the nowcast (*N*). Although *NFERS* can be a short timeseries whose autocorrelated errors is modelled using a standard first-order autoregressive (AR-1) formula, error propagation instead starts with the centered *observational* error (ε_I) and, by analogy with the impact of an assimilated observation, assumes a *symmetric* propagation through the gridded data:

$$
\begin{aligned}
\text{in situ} \quad I &= t + \epsilon_I \\
\text{nowcast} \quad N &= \alpha_N + \beta_N t + \lambda_N \epsilon_I + \epsilon_N \\
\text{forecast} \quad F &= \alpha_F + \beta_F t + \lambda_F(\lambda_N \epsilon_I + \epsilon_N) + \epsilon_F \\
\text{extended forecast} \quad E &= \alpha_E + \beta_E t + \lambda_E(\lambda_F(\lambda_N \epsilon_I + \epsilon_N) + \epsilon_F) + \epsilon_E \\
\text{revcast} \quad R &= \alpha_R + \beta_R t + \lambda_R(\lambda_N \epsilon_I + \epsilon_N) + \epsilon_R \\
\text{extended revcast} \quad S &= \alpha_S + \beta_S t + \lambda_S(\lambda_R(\lambda_N \epsilon_I + \epsilon_N) + \epsilon_R) + \epsilon_S
\end{aligned}
$$

So far, applications of this model have only resolved small true variance and a first order correlated error variance. The implication of a nonzero error correlation is that two datasets can be both physically independent and statistically dependent. The connection between small true variance and measurement model approximations that cannot be avoided, as well as the distinction between physical and statistical independence seem to require more attention.

4. Conclusions

Access to large geophysical datasets provides the freedom to extend existing measurement models by sampling in ways that might not have been tried before. A seemingly modest increase in number of samples, with a view to identify error autocorrelation in high resolution forecasts (but avoid error autocorrelation in a more sparse set of collocated observations) facilitates an identification of metrics of performance of both datasets and a linear calibration of one dataset to the other. Properties and solutions of this model, called INFERS, are currently being documented.

6. References

Bentamy, A., Piollé, J.-F., Grouazel, A., Danielson, R. E., Gulev, S. K., Paul, F., Azelmat, H., Mathieu, P.-P., von Schuckmann, K., Sathyendranath, S., Evers-King, H., Esau, I., Johannessen, J. A., Clayson, C. A., Pinker, R. T., Grodsky, S. A., Bourassa, M., Smith, S. R., Haines, K., Valdivieso, M., Merchant, C. J., Chapron, B., Anderson, A., Hollmann, R., Josey, S. A., 2017. Review and assessment of latent and sensible heat flux accuracy over global oceans. Remote Sens. Environ. 201, 196–218, doi:10.1016/j.rse.2017.08.016.

Danielson, R. E., Johannessen, J. A., Quartly, G. D., Rio, M.-H., Chapron, B., Collard, F., Donlon, C., 2018. Exploitation of error correlation in a large analysis validation: GlobCurrent case study. Remote Sens. Environ., submitted.

7. Acknowledgements

This work has been funded in part by the FP7 E-GEM and ESA GlobCurrent and Ocean Heat Flux projects.