Process-oriented models and statistical extraction of anthropogenic signals from measured environmental data

Anders Grimvall
Department of Mathematics, Linköping University, SE-58183 Linköping, Sweden.
E-mail: angri@mai.liu.se

Introduction

The currently used procedures to extract anthropogenic signals from measurements of the state of the environment are based on two main principles: statistical analysis of the variability of measured data, and process-oriented modelling of the environmental state under consideration. Several investigators have emphasised the statistical aspects of signal extraction or trend detection, and a variety of significance tests have been developed to test for the presence of persistent changes in measured data (e.g. Esterby, 1996). Other scientists have focused on the fact that a substantial fraction of the variation in environmental quality can often be attributed to natural fluctuations in weather conditions, and several regression-based procedures have been developed to remove weather-dependent fluctuations prior to the trend test (Thompson *et al*., 2001). Yet other scientists have employed physics-based models to derive how fluctuations in weather conditions can influence the studied environmental system. This article aims at bridging the gap between data-driven and process-oriented approaches to signal extraction. In particular, we show that different features of process-oriented deterministic models can be incorporated into statistical signal extraction techniques in a variety of different ways.

Principles of incorporating process-oriented models into statistical signal extraction techniques

Adopting the notation introduced by Grimvall *et al*. (2001) we let $\boldsymbol{y}$ denote a vector of observations of the state of the environment that are influenced by a vector $\boldsymbol{u}$ that represents natural fluctuations in e.g. weather conditions. Let us further assume that we have a process-oriented deterministic model that produces a unique response vector $\boldsymbol{y}^* = \boldsymbol{g}(\boldsymbol{u}, \boldsymbol{v})$ for any combination of the natural forcing $\boldsymbol{u}$ and the anthropogenic forcing $\boldsymbol{v}$ of the system under consideration. If the cited model produces correct predictions of $\boldsymbol{y}$, we can decompose $\boldsymbol{y}$ into one term

$$\boldsymbol{y}^* - E(\boldsymbol{y}^* \mid \boldsymbol{v}) = \boldsymbol{g}(\boldsymbol{u},\boldsymbol{v}) - \int \boldsymbol{g}(\boldsymbol{u},\boldsymbol{v})\mu(d\boldsymbol{u})$$

representing the modelled natural fluctuations for a given anthropogenic forcing $\boldsymbol{v}$, and another term

$$\boldsymbol{y} - \boldsymbol{y}^* + E(\boldsymbol{y}^* \mid \boldsymbol{v})$$

that may be interpreted as the estimated human impact on $\boldsymbol{y}$ of the anthropogenic forcing $\boldsymbol{v}$. Moreover, we can estimate $E(\boldsymbol{y}^* \mid \boldsymbol{v})$ by running the model with a representative set of input vectors $\boldsymbol{u}$ and computing the mean output $\bar{\boldsymbol{y}}$. If the model predictions are less perfect, we need other procedures to incorporate deterministic models into different types of signal extraction techniques. Here, we shall discuss the following options:

- the model predictions $\boldsymbol{y}^*$ are used as inputs to statistical signal extraction procedures;

- the process-oriented model $\boldsymbol{y^*} = \boldsymbol{g}(\boldsymbol{u}, \boldsymbol{v})$ is used to derive a sparsely parameterised meta-model that is subsequently incorporated into a statistical signal extraction procedure.

Signal extraction involving outputs of process-oriented models

This group of techniques involves several regression-based procedures. Instead of directly subtracting $\boldsymbol{y^*} - E(\boldsymbol{y^*} \mid \boldsymbol{v})$ from the vector of observed values, we can first transform the model outputs and then compute residuals. For example, we can form differences

$$e_i = y_i - b_0 - b_1(y_i^* - E(y_i^* \mid \boldsymbol{v})), \quad i = 1, \ldots, n,$$

between observed values and a suitable linear expression in the model outputs. More generally, we can form residuals

$$e_i = y_i - f(\boldsymbol{b}; y_i^* - E(y_i^* \mid \boldsymbol{v})), \quad i = 1, \ldots, n,$$

where f may be a nonlinear function and $\boldsymbol{b}$ is a vector of parameters.

If the timing of the model outputs is incorrect we may reduce the variance of the residuals by using the original model outputs as inputs to a transfer function. Letting B denote the backshift operator, this implies that we set

$$\omega(B) = \omega_0 - \omega_1 B - \ldots - \omega_s B^s,$$

$$\delta(B) = 1 - \delta_1 B - \ldots - \delta_r B^r,$$

and compute

$$\eta_i = y_i - \frac{\omega(B)}{\delta(B)} B^k (y_i^* - E(y_i^* \mid \boldsymbol{v})), \quad i = 1, \ldots, n,$$

where k denotes the delay between changes in output and output of the transfer function.

Furthermore, model outputs can be used to introduce so-called external drift in geostatistical models. If $Z(\boldsymbol{x},t)$ and $Z^*(\boldsymbol{x},t)$ respectively denote the actual and modelled state of the environment at location $\boldsymbol{x}$ and time t, we may write

$$Z(\boldsymbol{x},t) = b_0 + b_1 Z^*(\boldsymbol{x},t) + U(\boldsymbol{x},t)$$

where the residuals $U(\boldsymbol{x}, t)$ is a random function in space and time. Statistically this is regression model with correlated residuals. From a geostatistical perspective this equation implies that the drift of $Z(\boldsymbol{x}, t)$ is defined externally through an auxiliary variable $Z^*(\boldsymbol{x}, t)$ rather than from some smooth version of $Z(\boldsymbol{x}, t)$ itself. We may also regard the cited equation as a universal kriging decomposition into drift and fluctuation in which the basis drift functions are $(1, Z^*(\boldsymbol{x}, t))$ instead of monomials in the geographic coordinates $\boldsymbol{x}$.

Signal extraction involving meta-models

Process-oriented deterministic models do often focus on the short-term dynamics of the state of the environment, whereas extraction of anthropogenic signals is a matter of detecting long-term changes. Because of this, it may be advantageous to replace the original model with a simpler model, in which factors that mainly influence the short-term dynamics of the output are omitted. In particular, we can derive a meta-model, i.e. a statistical input-output model, by feeding the original model with a large number of artificially generated inputs and regressing the long-term model output on the inputs. What features of the process-oriented model that shall be retained in the meta-model can vary strongly with the predictivity of the original model. In this article, we

show how a process-oriented model of the leaching of nitrogen from agricultural soils can be used to derive a variety of more or less sparsely parameterised meta-models of the impact of different meteorological variables on the nitrogen leaching.

References

Esterby, S. (1996). Review of methods for the detection and estimation of trends with emphasis on water quality applications. *Hydrological Processes* **10**:127-149.

Grimvall, A., Wackernagel, H., Lajaunie, C. (2001). Normalisation of environmental quality data. In: Hilty, L.M., Gilgen, P.W. (eds) "*Sustainability in the Information Society*", pp. 581-590, Metropolis-Verlag, Marburg.

Thompson, M.L., Reynolds, J., Cok, L.H., Guttorp, P., Sampson, P.D. (2001). A review of statistical methods for the meteorological adjustment of ozone. *Atmos. Environ*. **35**:617-630.