Course title: EOTIST Standard course

Course subject: Modelling

Teacher: Marta Magnani

LESSON SM1.3
HOW TO BUILD A DATA-DRIVEN MODEL

EOTIST project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 952111

[H2020 WIDESPREAD-05-2020 (Twinning)]

TABLE OF CONTENTS

1. INTRODUCTION

An exhaustive knowledge of many natural processes and their drivers is still missing. Empirical models are a useful tool in the pioneering exploration of processes because they rely on the possibility to obtain information directly from the data without, or with just a few, assumptions.

Data-driven models are based on empirical relationships between measured variables. The aim of data-driven models is to represent a variable as a function of other measured variables, called predictors or explanatory variables. Thus, the modelled variable ($\widehat{Y}_t$, i.e. the variable that one wants to predict), at time t can be expressed as a function (F) of: time (t), other measured variables at time t (e.g. X_t and Z_t), the variables at a past time t-k (e.g. X_{t-k} and Z_{t-k}) and the same modelled variable at a past time t-k (i.e. Y_{t-k}, in that case the model is autoregressive). A general formulation of a data-driven model is the following:

$$\widehat{Y}_t = F(t, X_t, X_{t-k}, \dots, Z_t, Z_{t-k}, \dots, Y_{t-k}, \dots). \quad (1)$$

Different model formulations, namely different functional forms, as well as different candidate predictors are usually tested when building a model. The selection of the variables is performed by using statistical criterions: the predictors should not be cross-correlated (which can be excluded with a partial correlation analysis), the regression should be significant and the modelled variable ($\widehat{Y}_t$) should be as close as possible to the measured one (Y_t). The ability of the model to reproduce the measured variable is expressed by the explained variance (i.e. the ratio of the variances of the modeled and measured variable, $\sigma^2_{\widehat{Y}_t}/\sigma^2_{Y_t}$) or by the R^2. Hence, the optimal model is the one displaying the highest explained variance (or R^2). Moreover, when dealing with multi regression models (i.e. models including more than one predictor), the inclusion of more predictors may increase the explained variance of the model to the detriment of the model parsimony. A further constrain on the model can be obtained with the Akaike Information Criterion (AIC). The AIC measures the goodness of a statistical model based on a trade-off between the residual variance (that is, the variance of the residuals, $\widehat{Y}_t - Y_t$) and model parsimony (that is, with a penalty proportional to the number of free parameters). By this criterion, the empirical model having the lowest AIC should be preferred. Such criterion is used to compare models that differ in their functional form and/or in the accounted predictors.

As a final remark, predictors are not necessarily drivers of the process, thus the nexus of causality between the predictor and the modelled variable need to be tested thereafter. In the following, such general descriptions are applied to build models of carbon dioxide (CO_2) fluxes in both the high-altitude Alpine and the high-Arctic ecosystems.

QUANTITATIVE MODELS FOR TERRESTRIAL CO_2 FLUXES

Over lands, the net CO_2 flux is made of two components: photosynthetic uptake by plants, and emissions by (plant) autotrophic and (microbial) heterotrophic metabolic activity. The plant photosynthetic uptake is called the gross primary production (GPP, which assumes negative values by convention) and the total emission is called ecosystem respiration (ER, which assumes positive values).

The air (or soil) temperature and the photosynthetic active radiation are well known drivers of ER and GPP, respectively. In detail, an exponential dependence of ER on the environmental temperature and a response of GPP to light intensity through a Michaelis-Menten function are commonly assumed:

$$ER = a\, e^{bT}, \quad (2)$$

$$GPP = \frac{F\alpha\, R}{F+\alpha\, R}. \quad (3)$$

Here, the temperature *T* is expressed in Celsius, a is a free parameter, corresponding to the respiration at 0 °C, *b* is the temperature sensitivity of respiration, *R* is the incoming solar radiation, *F* is the maximum photosynthetic flux for infinite light supply and α is the apparent quantum yield, i.e. the photosynthetic response at low light level. Reference to the original papers are in the attached slides.

Field studies suggest that Eq. (2-3) reproduce the flux variations only partially in both the Alpine and the Arctic ecosystems. Hence, multi regression models are usually built to identify additional drivers. Since the regression parameter in Eq. (2-3), namely *a, b, F* and α, were shown to depend on other factors, the regression parameters of the classical functions can be thought as functions of additional drivers (e.g. $a = a(x_1, x_2, ...)$ and $b = b(x_1, x_2, ...)$, with x_1, x_2 being any other measured variable). The function of the additional drivers, which are not known a priori, can be approximated with polynomial series. If the additional drivers produce small changes in the regression parameters, then Taylor expansions are possible. Then, multi regression models were built by retaining the first order of the expansion and testing all the measured variables as candidate additional predictors See also Magnani et al (2020).

CO_2 fluxes and basic meteoclimatic variables (including air temperature and moisture, solar irradiance, atmospheric pressure, and soil temperature and moisture) were measured at both an Alpine and an Arctic site. In the Arctic site, located in the Svalbard Islands (Norway), data-drive models were used to study the factors influencing the fluxes at different spatial scales (from about 20 cm to 100 m), as well as the differences between the fluxes measured for different vegetation types. This was achieved by comparing regression parameters of the models estimated for different classes of vegetation, thus assessing whether, for different vegetation types, the fluxes display a different response to the same values of the predictors. In the Alpine study, located in the Gran Paradiso National Park (Italy), three sites characterized by different parental material were identified within the same watershed to analyze the possible site-to-site differences within the same environment. Measurement designs are detailed in the slides.

In the Arctic site, the temporal variability of the fluxes at a fixed point along 24 hours was well reproduced by the classical drivers: the air temperature for ER (Eq. 2), and the solar irradiance for GPP (Eq. 3). However, when zooming out to the site scale, only a small part of flux variability was explained by the classical drivers and multi regression models were needed. At the site scale, the measured meteoclimatic variables, together with the hour and day of sampling were used to build multi regression models. The spatial and temporal patterns of the fluxes were reproduced by the following equations:

$$ER = (a_0 + a_1 VWC + a_3\, GFC)\, e^{b_0 T_a}, \quad (4)$$

$$GPP = \frac{F_0 \alpha_0\, R}{F_0+\alpha_0\, R}\, (A_0 + A_1 VWC + A_2\, GFC)\ , \quad (5)$$

Where *VWC* is the soil moisture (volumetric water content), *GFC* the green fractional cover (between 0 and 1), T_a is the air temperature and *R* is the total incident solar irradiance. The models were then parameterized for both different classes of vegetation (vascular, non-vascular and mixed vegetation) and different species, representing the most abundant vascular species in the site: Carex spp., Dryas octopetala, Salix polaris, Saxifraga oppositifolia and Silene acaulis. Significant differences between vegetation classes and between species were observed in the flux dependence on the *GFC* (i.e., significantly different a_1 and A_1 parameters). The original paper is attached.

In the Alpine case, for each of the three measurement sites, the selected the models where:

$$ER = (a_0 + a_1 VWC + a_2\, Pr + a_3\, DOY)\, e^{b_0 T_a}, \quad (6)$$

$$GPP = \frac{F_0 \alpha_0\, R}{F_0 + \alpha_0\, R}\, (A_0 + A_1 VWC + A_2\, DOY)\ , \quad (7)$$

where the same conventions of Eq. (4-5) were used, and *Pr* is the atmospheric pressure and *DOY* the day of the year (from 1 to 365). Moreover, comparing the regression parameters of Eq. (6-7) estimated for different sites, significant differences were observed in the flux dependence on *VWC*, namely A_1 and a_1, possibly suggesting an additional influence of the parental material, via soli characteristics, on the fluxes. The original paper is attached.

Hence, in both the Alpine and the Arctic case study the soil moisture (*VWC*) and a vegetation descriptor, namely the *GFC* in the Arctic and *DOY* that was interpreted as a proxy of phenology in the Alps, were identified as additional predictors of the fluxes beyond the classical drivers. Further comments are present in the slides.

The slides are complemented by the original papers.