Funded by
European Union

Grant Agreement 952111
H2020 WIDESPREAD-05-2020 (Twinning)

Standard course – Modelling

Lesson SM1.3 – How to build a data-driven model

Marta Magnani – Institute of Geoscience and Earth Resources (CNR, Italy)

Data-driven models

Empirical models are a useful tool in the pioneering exploration of processes. Data-driven models are based on empirical relationships between measured (or estimated) variables and aim at representing a variable (*Y*) as a function of other measured variables (X_i, *i=1,2,3,…*), called predictors or explanatory variables. The steps to build a data-driven model are:

1. Identify different model formulations, possibly inspired by literature or by the characteristics of data.

1. Test different candidate predictors

1. Use statistical criterions to select the best model:
   - ✓ the predictors should not be cross-correlated (excluded by a partial correlation analysis)
   - ✓ the regression should be significant (i.e. P-value of predictors lower than a chosen threshold of significane)
   - ✓ the modelled variable () should be as close as possible to the measured one (Y_t).
   - ✓ the model should be efficient (i.e. it maximize representativeness while minimizing the number of parameters), for instance minimizing the Akaike Information Criterion (Akaike, 1974; Burnham and Anderson, 2002)

QUANTITATIVE MODELS FOR TERRESTRIAL CO_2 FLUXES

Measurements at the Alpine site

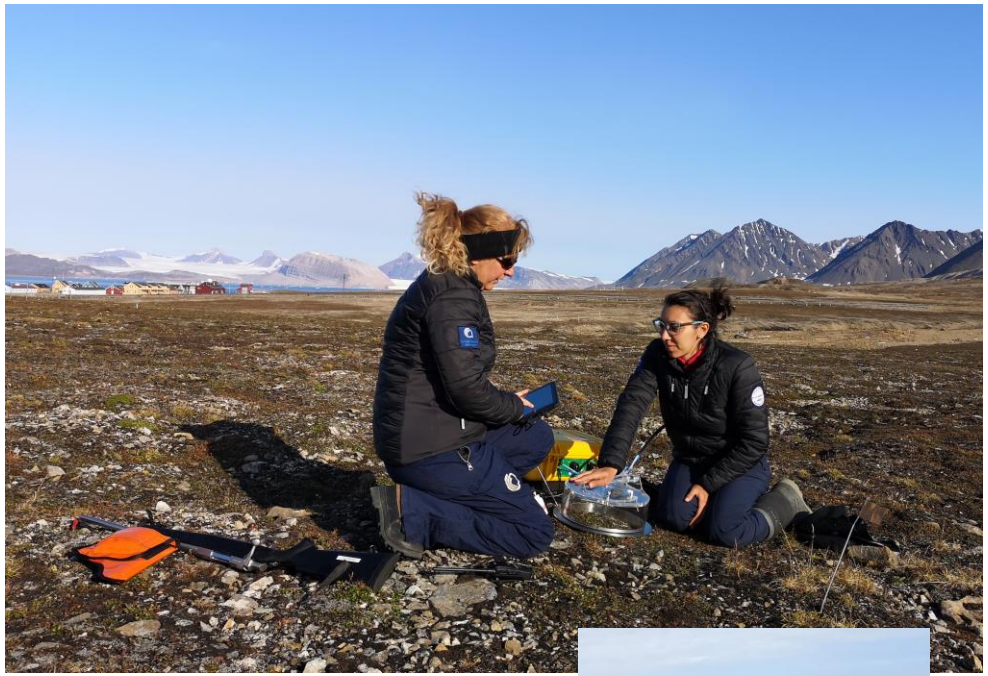
Artic measurement site at Bayelva, Spitsbergen, Norway. In situ samplings of:

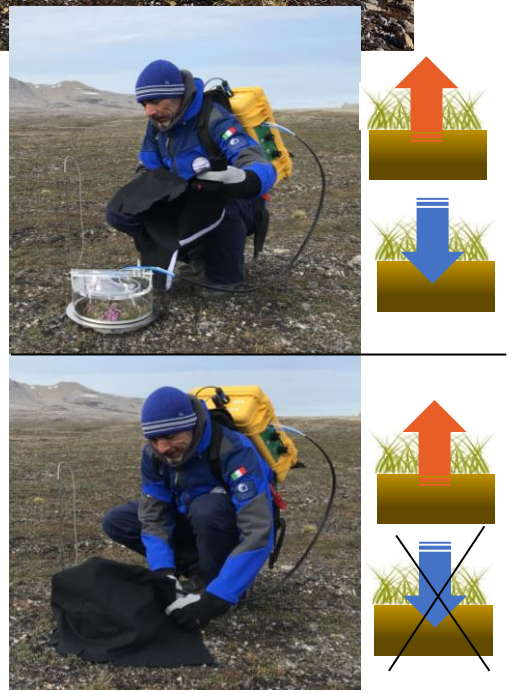
1. CO_2 fluxes with portable accumulation chambers: net flux (NEE) & emissions (ER) → uptake (GPP=NEE-ER)

1. Meteoclimatic variables (Air pressure, Air temperature, air moisture, solar irradiance, soil temperature, soil moisture)

1. Vegetation green fractional cover

1. Vegetation classification
   - **Vascular vs non-vascular**
   - **Different vascular species**

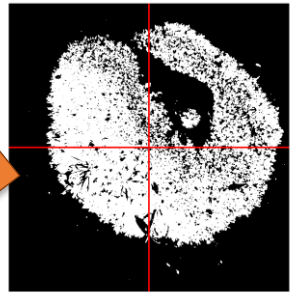




1. Flux measurement NEE & ER

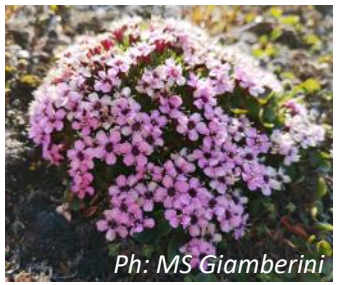
3. Green Fractional Cover

method: Liu & Pattey, 2010



4. Vegetation classification

Dryas octopetala



Silene acaulis

Experimental design

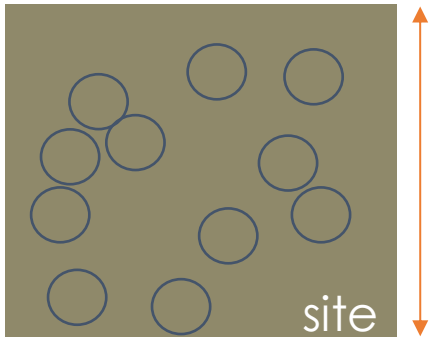
Flux temporal variability

Measurements performed at a fixed location (i.e. point) for 24 hours every 2 hours.

Point scale (0cm)

Flux spatiotemporal variability

fluxes measured at different points randomly distributed over the site. Measurements repeated in consecutive days



Site scale (100m)

Models of the Artic fluxes

Well-Known drivers

Temporal variability

$$EcosystemRespiration = ae^{bTemperature}$$

$$GrossPrimaryProduction = \frac{F\alpha Radiation}{F + \alpha Radiation}$$

Point scale (0cm)

Spatiotemporal variability

+ ???

+ ???

Site scale (100m)

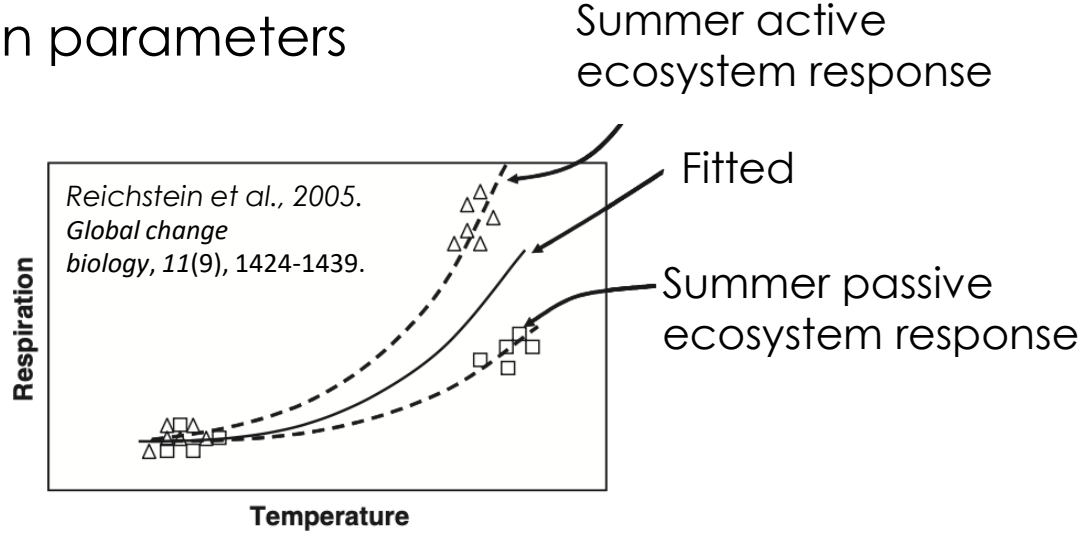
At a fixed point along 24h: the temporal variability of fluxes was well reproduced by the classical drivers, temperature for ER and solar irradiance for GPP.

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.

Additional variables

Idea: additional drivers modify the regression parameters
→ Taylor expansions

$$ER = a(x_1, x_2)e^{b(x_1,x_2)T}$$
$$GPP = \frac{F(y_1, y_2)\alpha(y_1, y_2)R}{F(y_1, y_2) + \alpha(y_1, y_2)R}$$



Where and are chosen among the measured variables. Statistical selection of both model and variables. The model thus obtained is:

$$\textit{emissions} \begin{cases} ER = (a_0 + a_1 VWC + a_2 GFC)e^{b_0 T} + \varepsilon \\ \textit{uptake} \end{cases} \quad \left| GPP = \frac{F_0 \alpha_0 R}{F_0 + \alpha_0 R}(A_0 + A_1 VWC + A_2 GFC) + \varepsilon \right|$$

T=temperature
VWC=soil moisture
GFC=Green Fractional Cover
R=solar irradiance

Multi regression models

Well-Known drivers

$$\boldsymbol{EcosystemRespiration} = ae^{\boldsymbol{bTemperature}}$$

Lloyd & Tylor, 1994. Funct. Ecol. 315–323.

$$\boldsymbol{GrossPrimaryProduction} = \frac{F\alpha\boldsymbol{Radiation}}{F + \alpha\boldsymbol{Radiation}}$$

Ruimy et al., 1995. Adv. Ecol. Res. 26, 1–68

Inserting additional descriptors in multi regression models

$$\boldsymbol{ER} = (a| \quad |0 + a_1\boldsymbol{GreenFractionalCover} + a_2\boldsymbol{soilmoisture})e^{\boldsymbol{bTemp}}$$

$$\boldsymbol{GPP} = (A| \quad |0 + A_1\boldsymbol{GreenFractionalCover} + A_2\boldsymbol{soilmoisture})\frac{F\alpha\boldsymbol{Rad}}{F + \alpha\boldsymbol{Rad}}$$

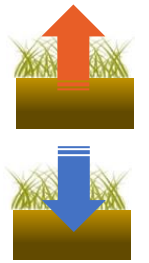
Ref to: Magnani et al., 2022. Scientific reports, 12(1), 1-14.

In both the Alpine and the Arctic case study the soil moisture (*VWC*) and a vegetation descriptor, namely the *GFC* here 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.

Role of different plant types

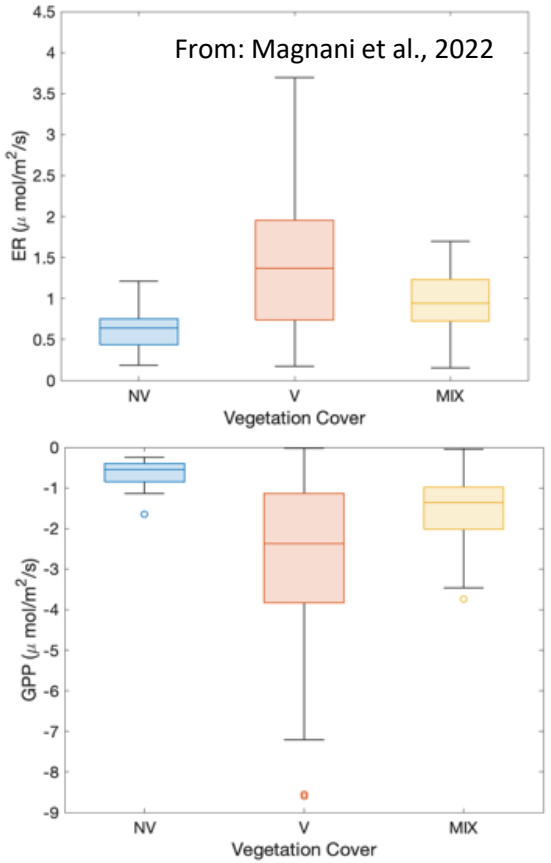
$$\boldsymbol{ER} = (a_0 + a_1\boldsymbol{GreenFractionalCover} + a_2\boldsymbol{soilmoisture})e^{b\boldsymbol{Temp}}$$

$$\boldsymbol{GPP} = (A| \quad |0 + A_1\boldsymbol{GreenFractionalCover} + A_2\boldsymbol{soilmoisture})\frac{F\alpha\boldsymbol{Rad}}{F + \alpha\boldsymbol{Rad}}$$



Non vascular

vascular



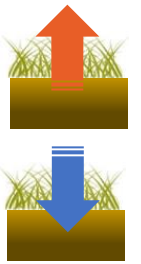
Exploring more in depth the role of different plan types, the dataset was divided into three subset according to the prevailing vegetation cover type observed at each sampling point. The three classes:
V: vascular vegetation (herbs and prostrated plants),
NV: non-vascular vegetation (likens, mosses and bacterial soil crust) MIX: mix of vascular and non-vascular vegetation (no clear prevalence).

Reparametrizing the models for these three classes, significant differences were observed in the flux dependence on the *GFC* (i.e. significantly different a_1 and A_1 parameters) between classes.
→ **different physiological functioning of vegetation groups**

Role of different species

$$\boldsymbol{ER} = (a_0 + a_1\boldsymbol{GreenFractionalCover} + a_2\boldsymbol{soilmoisture})e^{b\boldsymbol{Temp}}$$

$$\boldsymbol{GPP} = (A| \quad |0 + A_1\boldsymbol{GreenFractionalCover} + A_2\boldsymbol{soilmoisture})\frac{F\alpha\boldsymbol{Rad}}{F + \alpha\boldsymbol{Rad}}$$



Vascular class was the most variable in terms of green fractional cover → species-specific samplings.

The models were parameterized using the species-specific samplings and significant differences were observed in the flux dependence on the *GFC* between two groups of species: *SI-DR* vs *SL-SX*, while *CX* showed hybrid characteristics between these two groups of species.

→ **different physiological functioning of vegetation *species* groups**

5 classes of most abundant vascular plants identified

Silene acaulis

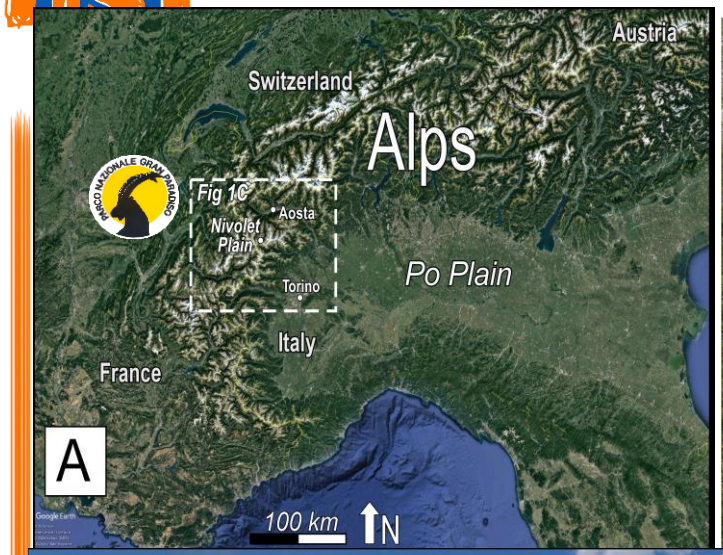
Carex Spp.

Salix polaris

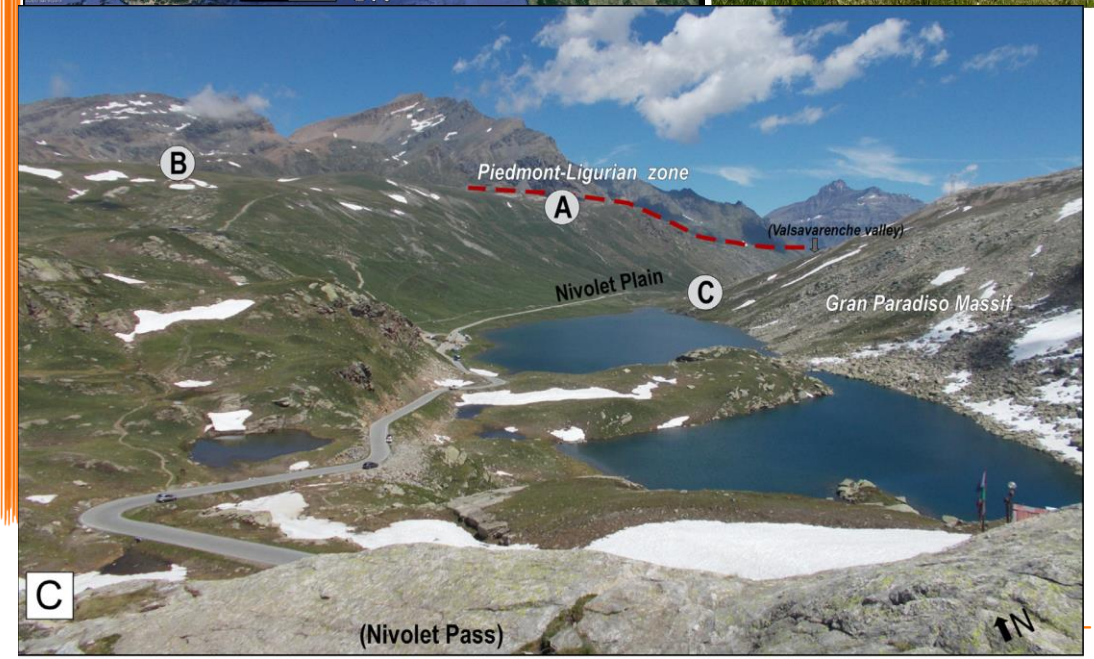
Dryas octopetala

Saxifraga oppositifolia

Measurements at the Alpine sites



Alpine measurements site at the Nivolet plain, in the Gran Paradiso National Park, western Italian Alps (45° 28′ 43″ N 7° 08′ 32″ E).



Measurements in summers 2017->2021

- Carbon dioxide fluxes
- Soil temperature and moisture
- Air temperature, moisture and pressure
- Solar irradiance
- Vegetation information from pictures

On **3 sampling plots** characterized by different parental material:

Siite A) carbonate rocks

Site B) glacial till

Site C) gneiss rock

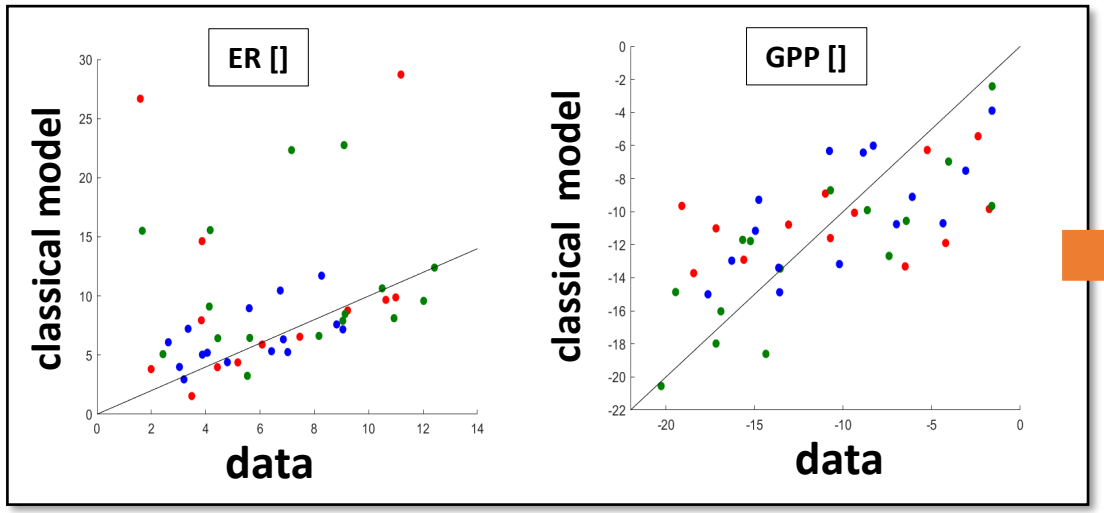
Multi regression models

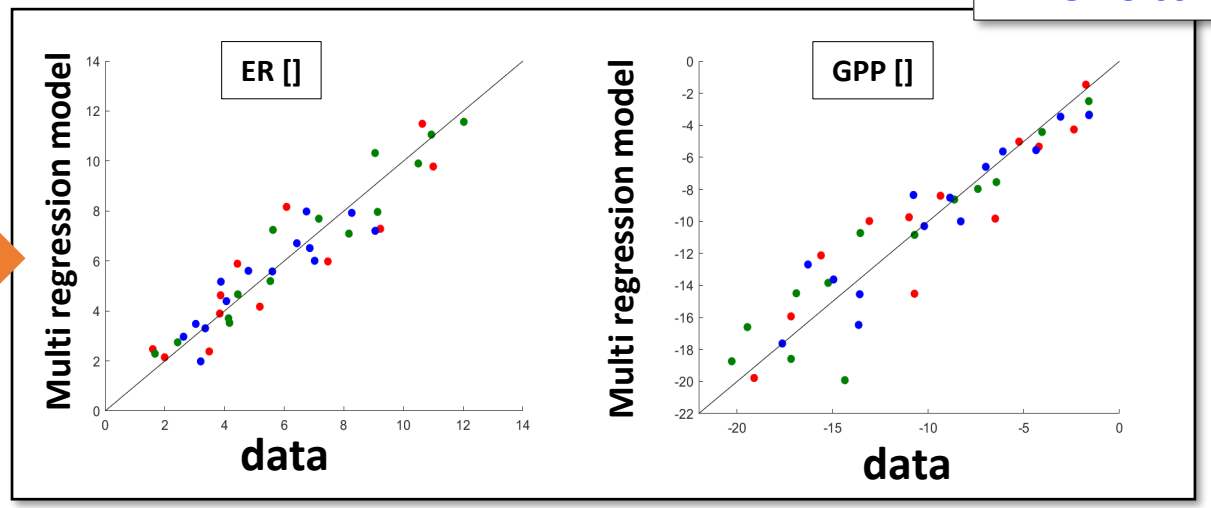
emissions $ER = ae^{bT}$

uptake $GPP = \frac{F\alpha R}{F + \alpha R}$

$$\begin{cases} ER = (a_0 + a_1 VWC + a_2 Pr + a_3 DOY)\, e^{b_0 T} \\ GPP = \frac{F_0 \alpha_0}{F_0 + \alpha} \end{cases}$$

- Carb.
- Glac.
- Gneiss





T=temperature; R=solar irradiance; VWC=soil moisture; Pr=air pressure; DOY = Day Of the Year

- **Soil moisture** limits both carbon emissions and uptake
- **Air pressure** may have a leaking effect on emissions
- Significant differences among parameters from different sites → **Soil characteristics** are additional constraints
- The **Day of the Year** explained a large part of fluxes variability → What it is a proxy for?

Ref to Magnani et al., STotEn, 2020