EC Data Postprocessing, sharing and more…

Dario Papale

University of Tuscia – Viterbo, Italy

ICOS - Ecosystem Thematic Center

darpap@unitus.it

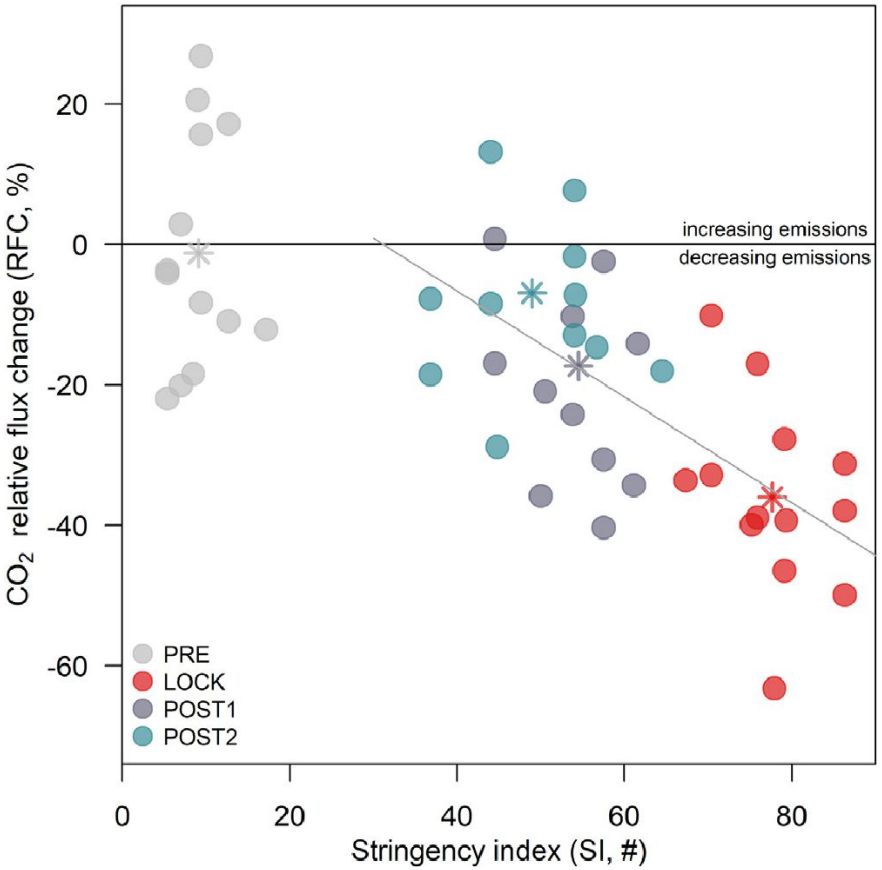
UNIVERSITÀ DEGLI STUDI DELLA Tuscia

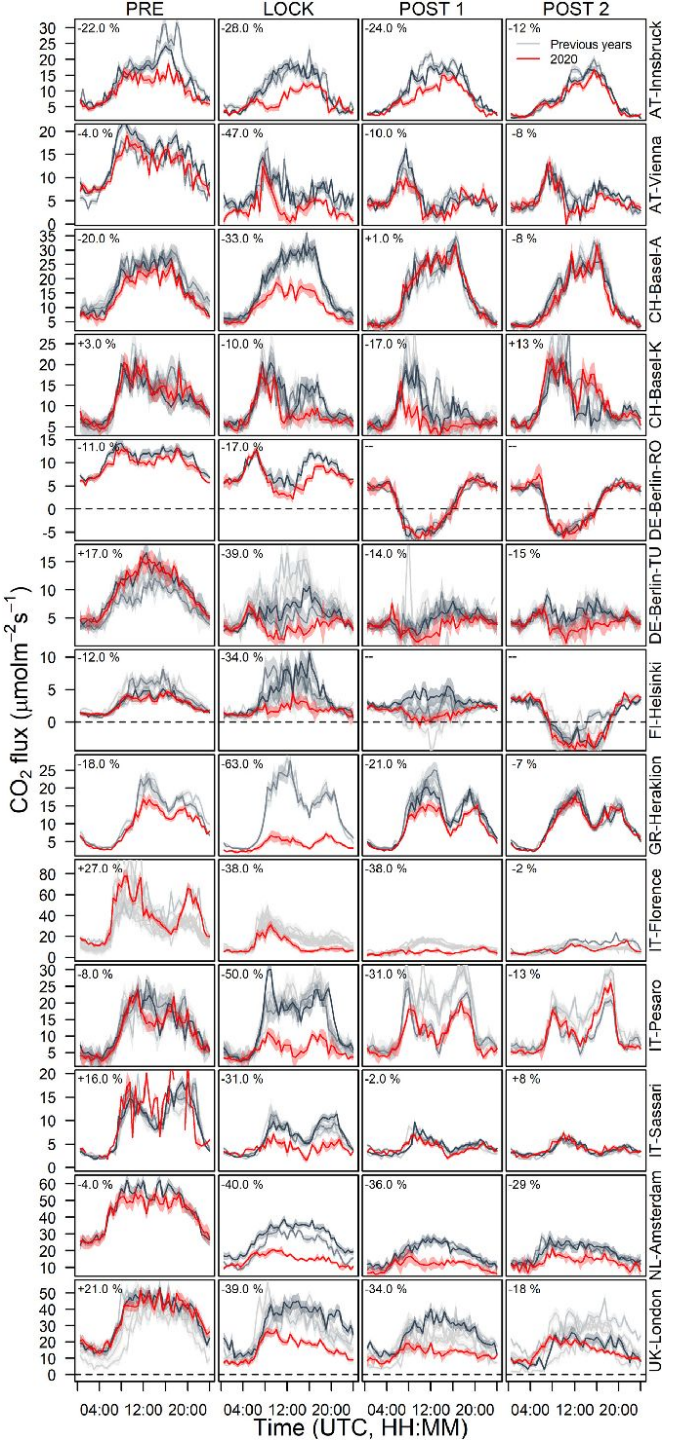
ICOS Ecosystem Thematic Centre

How to end up working on EC even if you do not want…

Did my master in Remote Sensing, never wanted to work on eddy covariance	362 ppm CO_2	1996
In my PhD I started to use EC data (really few sites)	375 ppm CO_2	2003
Realized that the data available were too heterogeneous, so together with Markus Reichstein started to work on standardization	379 ppm CO_2	2005
Created a first standardized collection of EC data (LaThuile Collection)	383 ppm CO_2	2007
Updated with new data and processing for the FLUXNET 2015 Collection	399 ppm CO_2	2015
Coordinator of the European ICOS network of EC sites	403 ppm CO_2	2016
I'm here	418 ppm CO_2	

EC can be applied over different surfaces (e.g. urban)





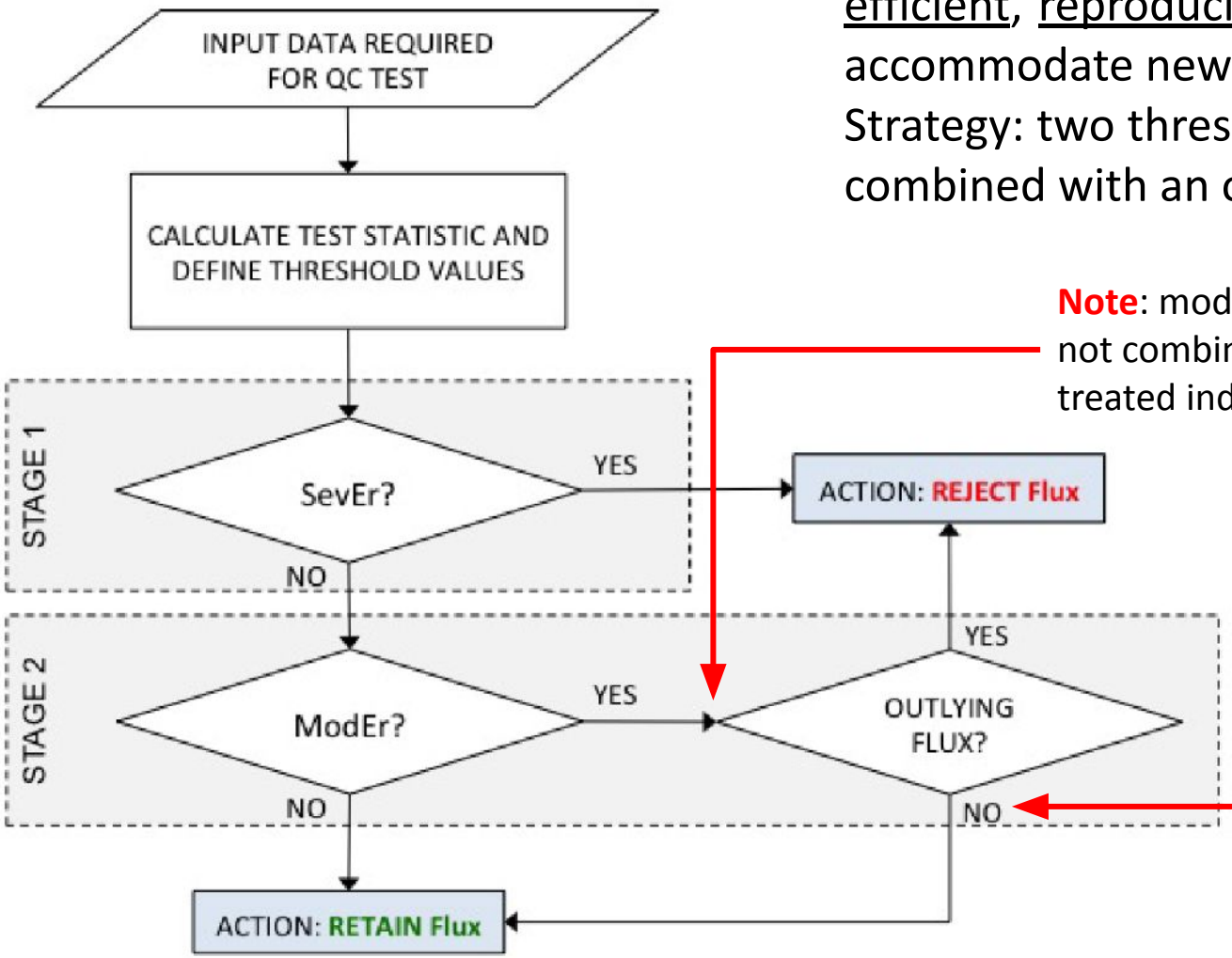
Nicolini et al. 2022

QA-QC filtering strategy

A good data filtering is critical for the quality of the final product. Objective QAQC are complex but needed, making maximum use of metadata, instruments flags, status indicators and statistical tests.

Types of error	**Test *(examples)***
Integrity of raw-data (gaps, diagnostics of the instruments, wind sectors etc.)	% of not available records
Instrumental problems not detected by the diagnostic e.g. - Signal resolution (limited digits) - Dropouts (continuous fix value) - Presence of spikes - Discontinuities (jumps in the	Statistical tests (e.g. kurtosis)
Violation of stationary conditions	Foken and Wichura (1996), Mahrt (1998)
Lack of well-developed turbulence conditions	Foken and Wichura (1996)
Suitability of spectral correction procedure	Spectral correction factor magnitude

QA-QC filtering strategy



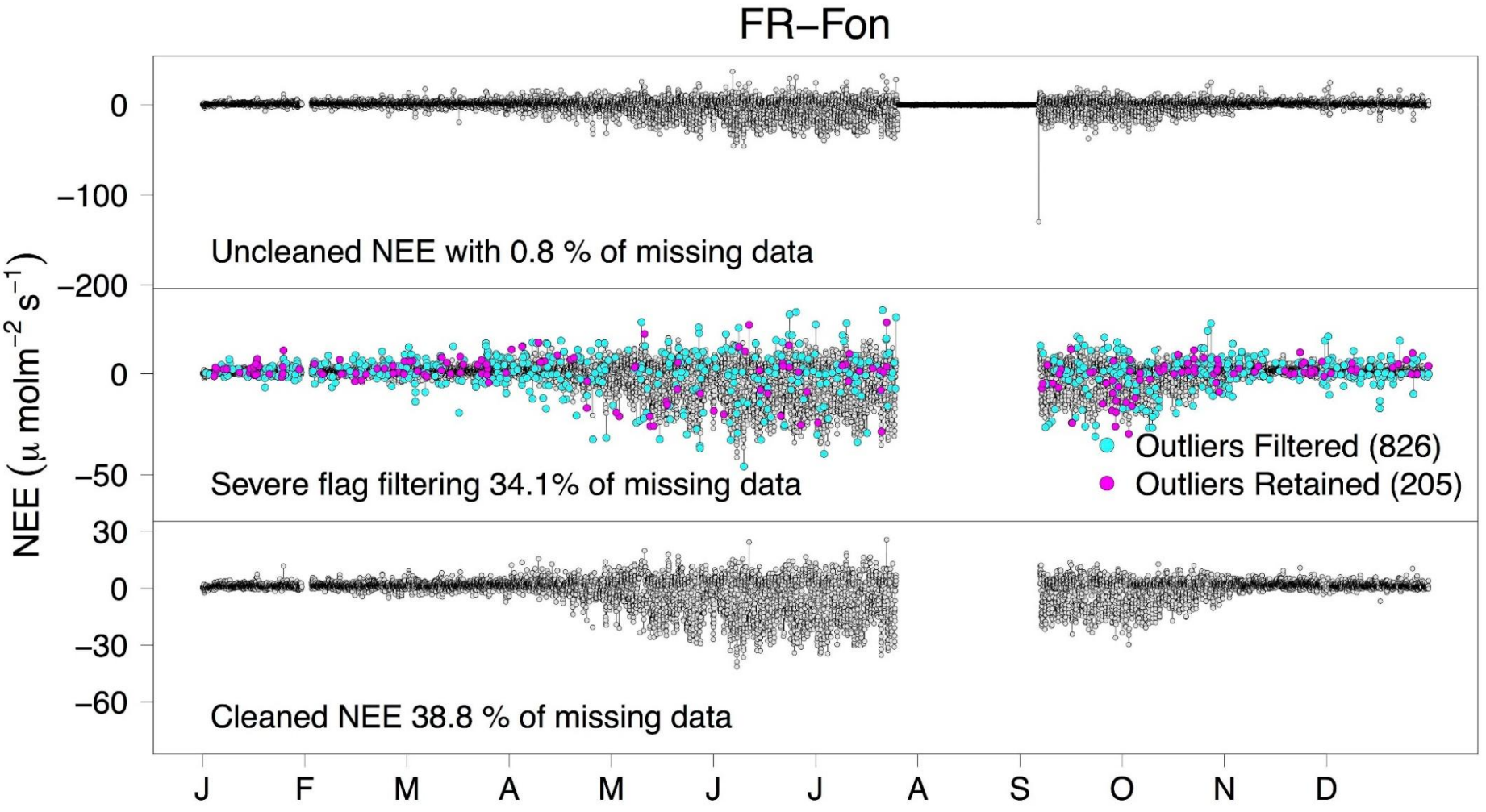
Data quality filtering should be efficient, reproducible and flexible to accommodate new tests.
Strategy: two thresholds for each test combined with an outlier detection.

Note: moderate flags are not combined. They are treated individually

Note: if outlier but without flags it is retained

Vitale et al., 2020

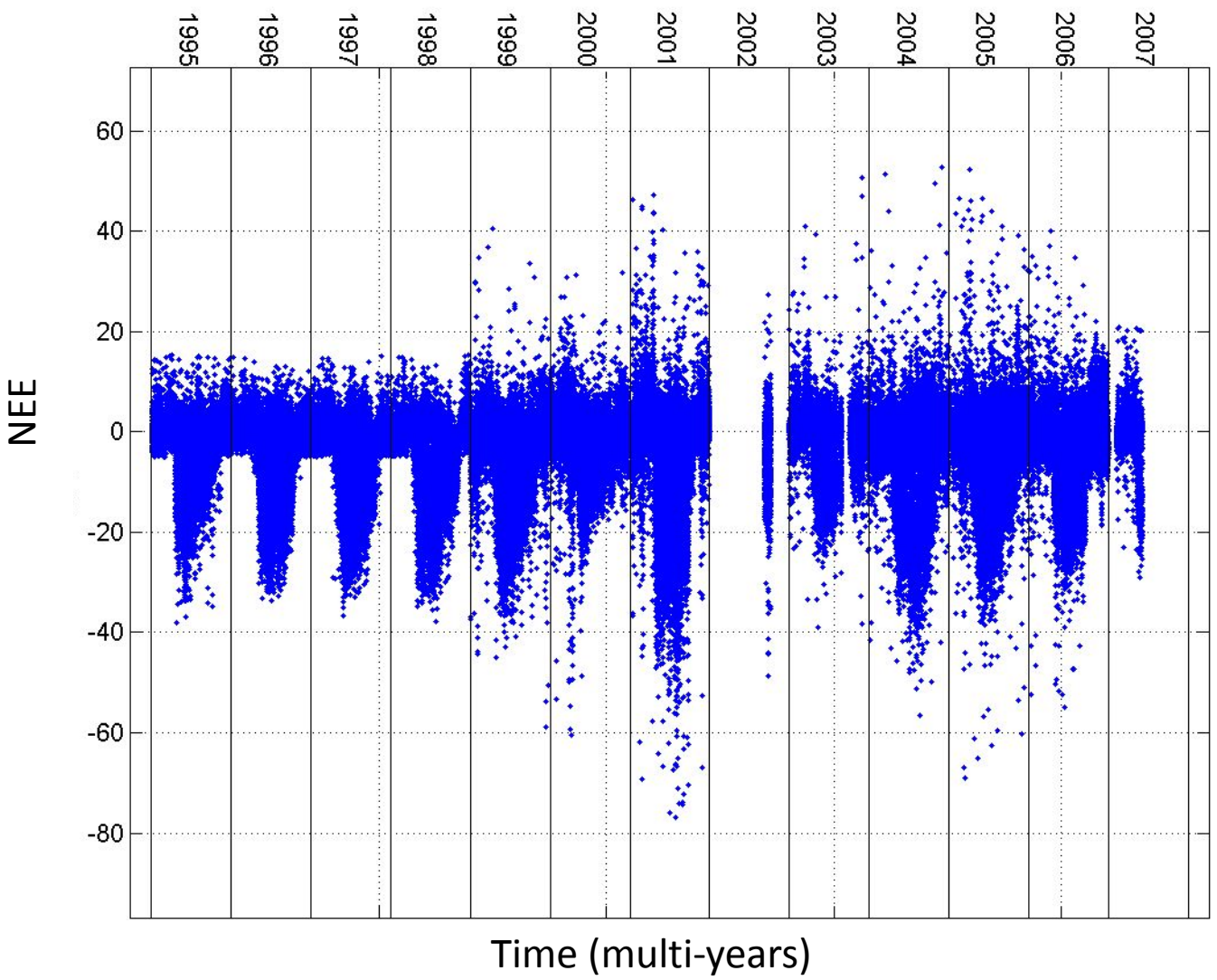
QA-QC filtering strategy

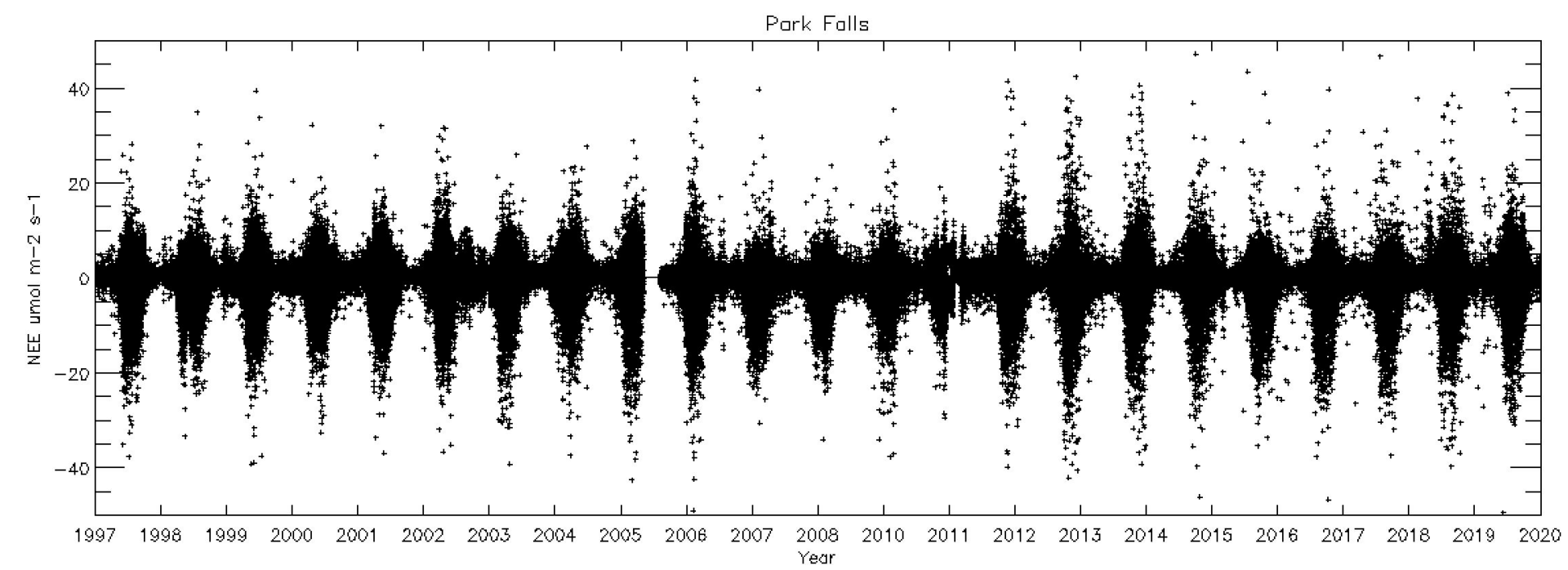


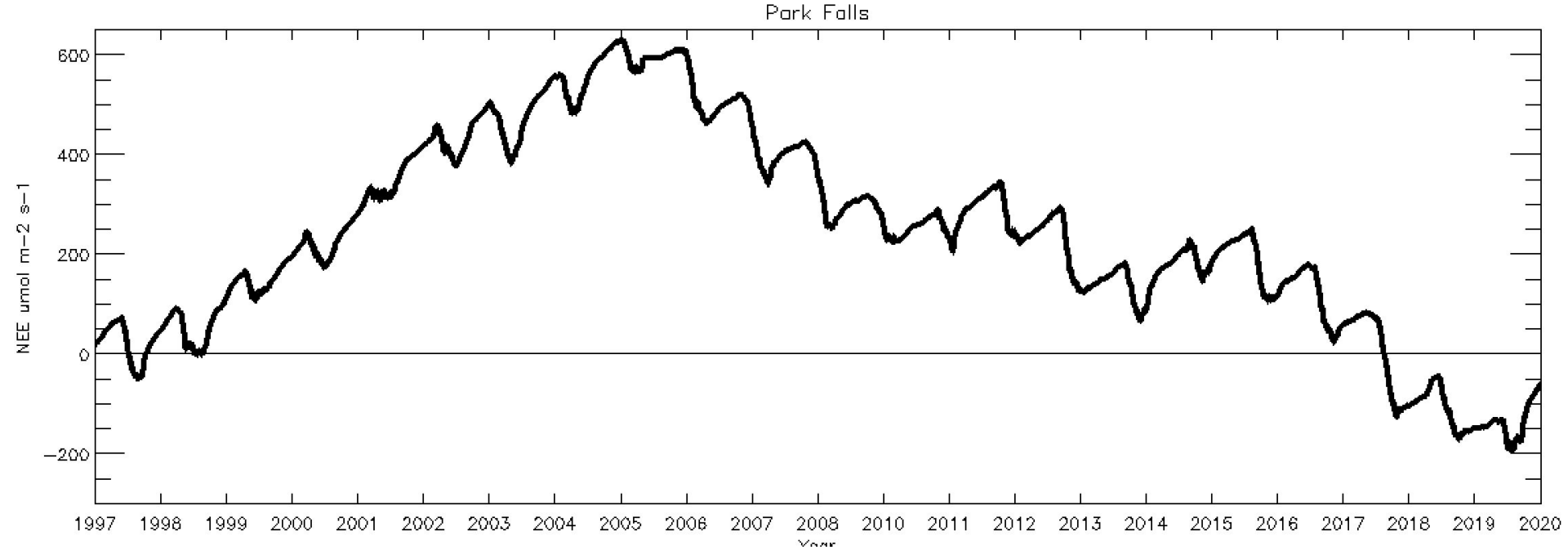
Vitale et al., 2020

Heterogeneity in a timeseries

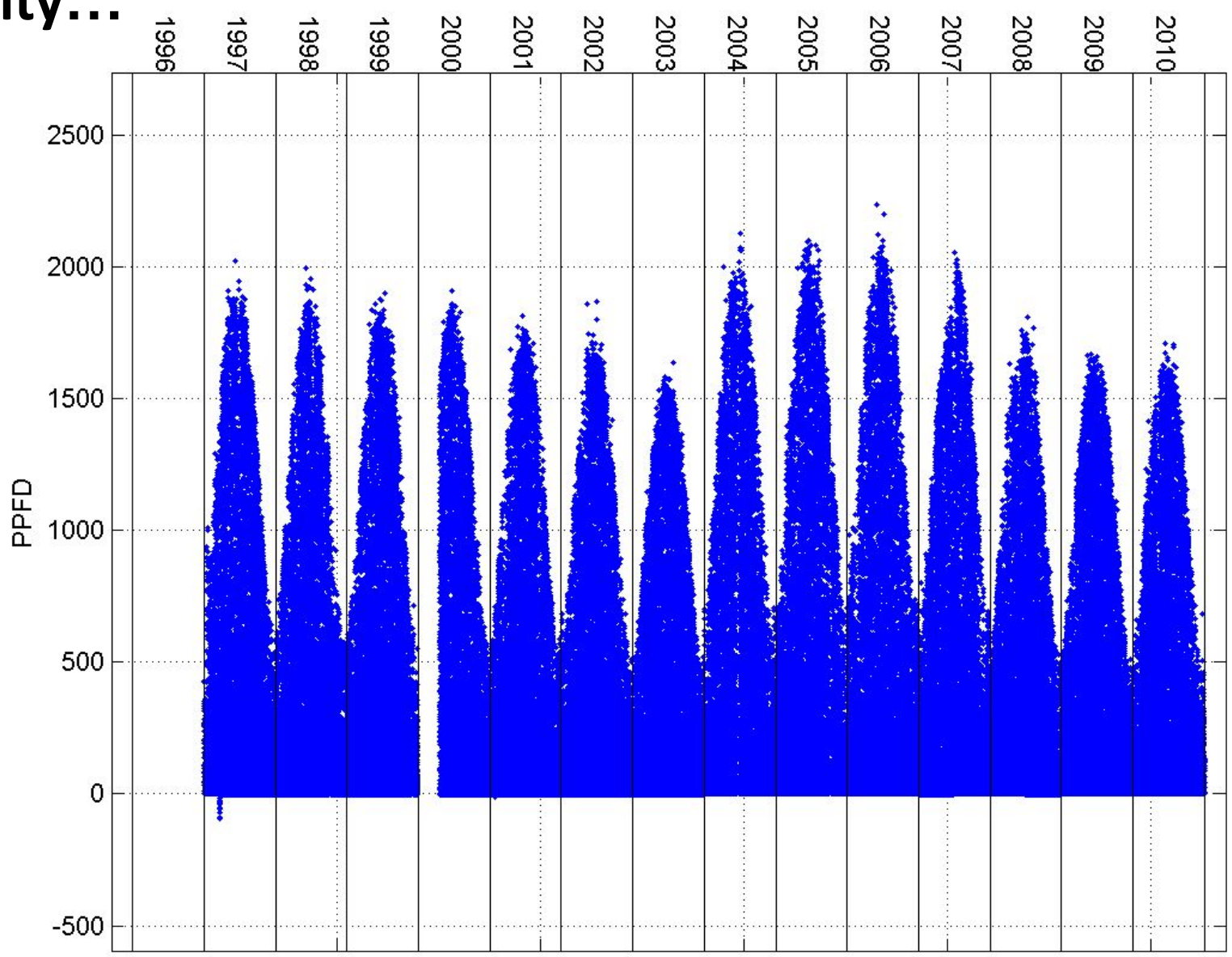
Real ecosystem variability? Change of sensors or setup? Different processing or QC?



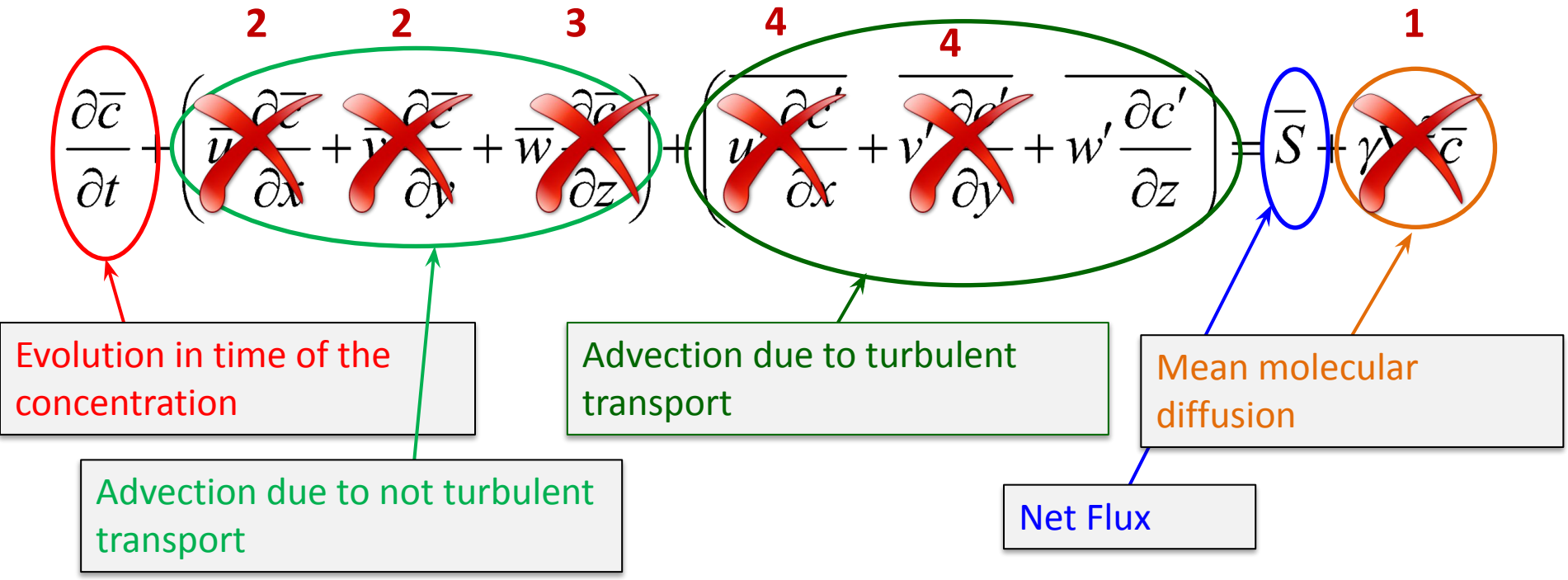

Park Falls
NEE umol m-2 s-1
40
20
0
-20
-40
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
Year


Park Falls
NEE umol m-2 s-1
600
400
200
0
-200
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020

Do not forget the meteo sensor quality…



The eddy covariance fluxes



1. Molecular diffusion is minor in turbulent transport regime
2. Horizontal variations of <u>mean</u> concentration can be neglected
3. <u>Mean</u> vertical velocity is almost zero
4. Turbulence is homogeneous in the different horizontal directions

Net Ecosystem Exchange calculation

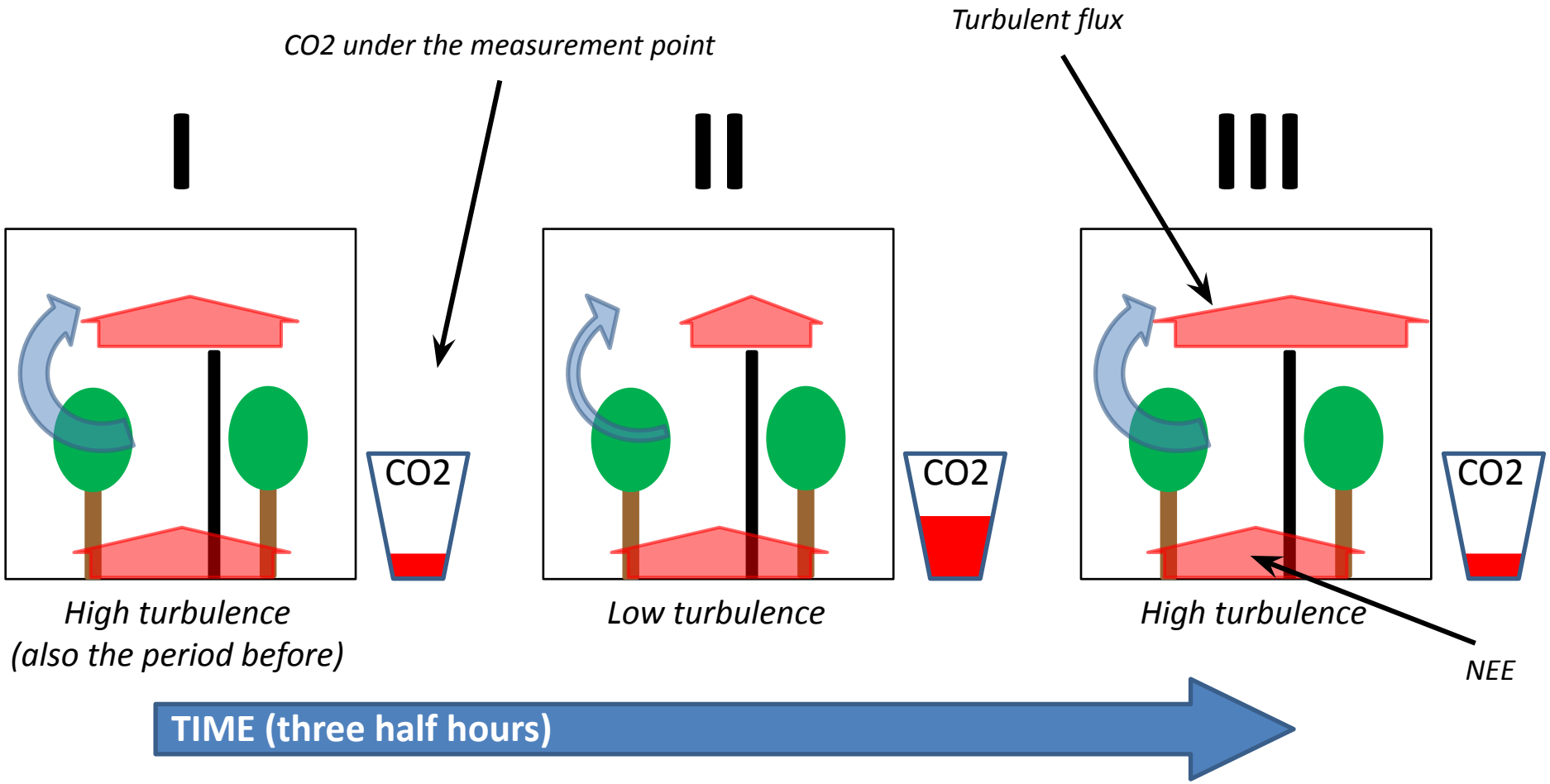
$$F_C = F_C^{TUR} + F_C^{STO} + F_C^{ADV}$$

Turbulent fluxes
Measured by the EC system

Storage
Measured using additional systems (profile)

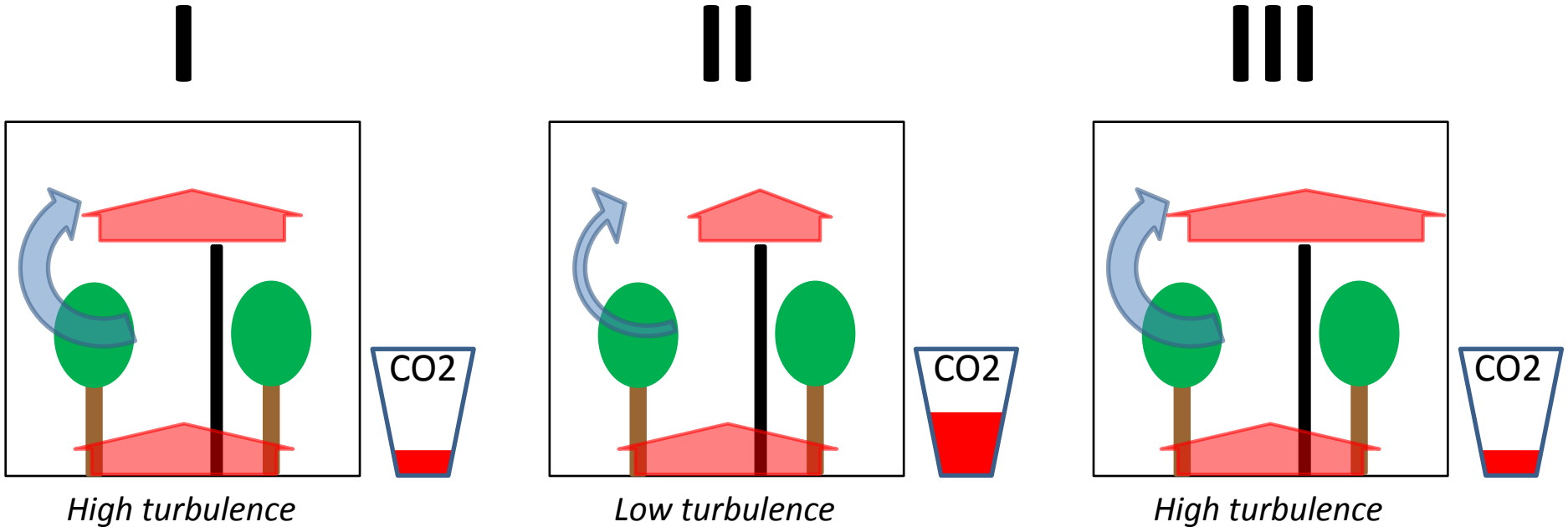
Advection
Difficult to measure
Corrections needed

STORAGE (Sc) example: nighttime, summer.



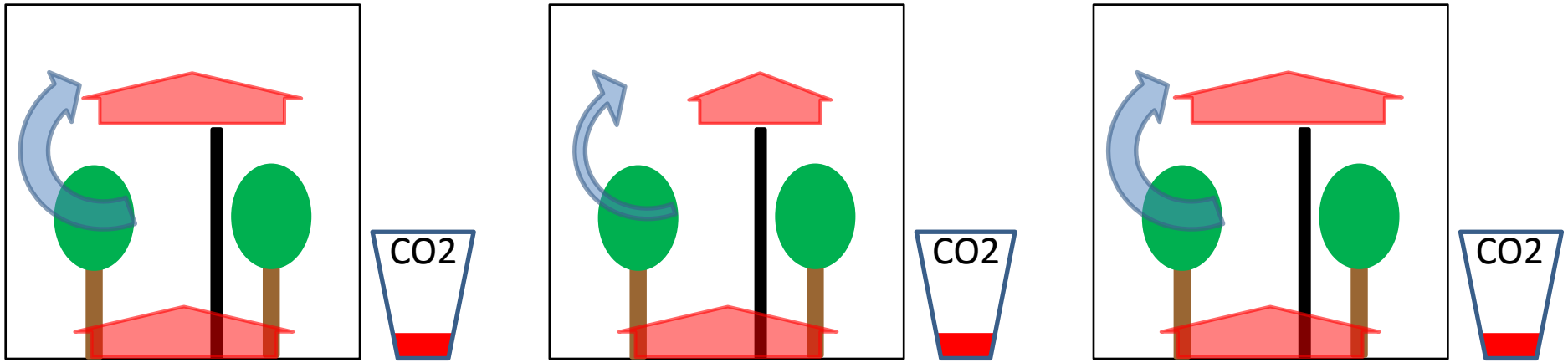
All the examples from here will be on CO2

STORAGE (Sc) example: nighttime, summer.



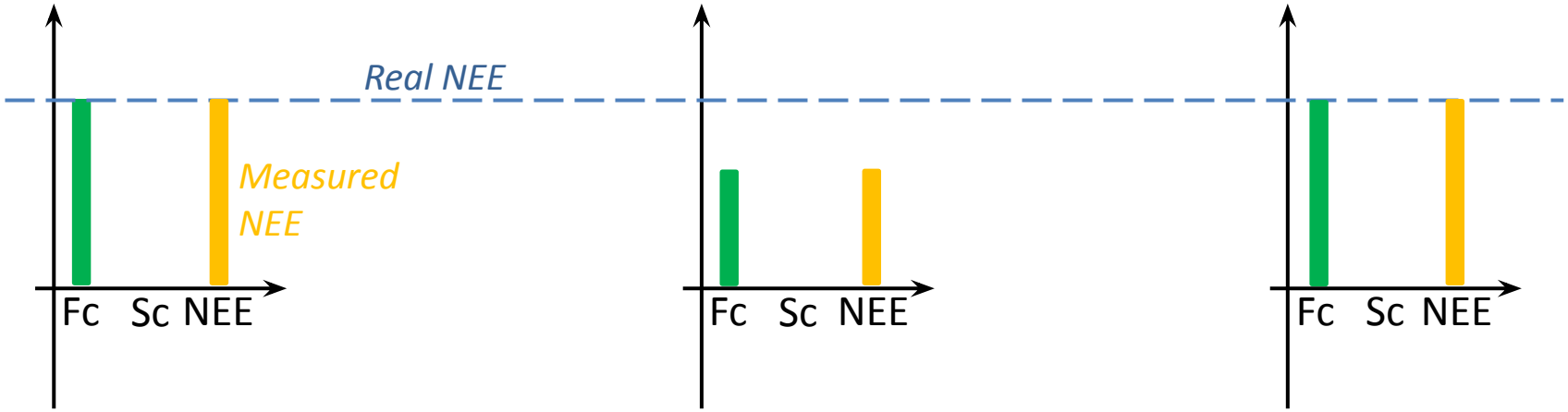
ADVECTION example: nighttime, summer.

I II III



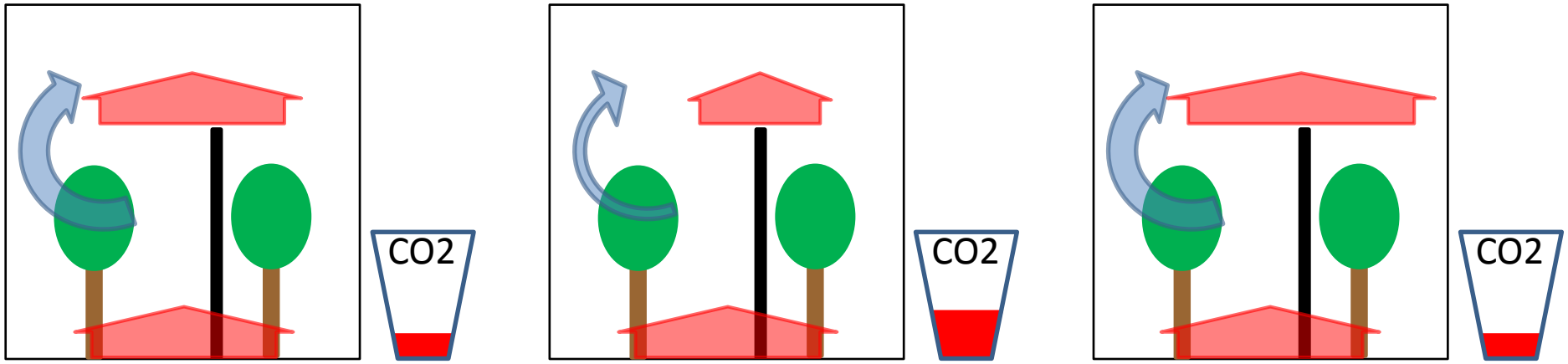
High turbulence (also before) | *Low turbulence* | *High turbulence*

TIME (three half hours)



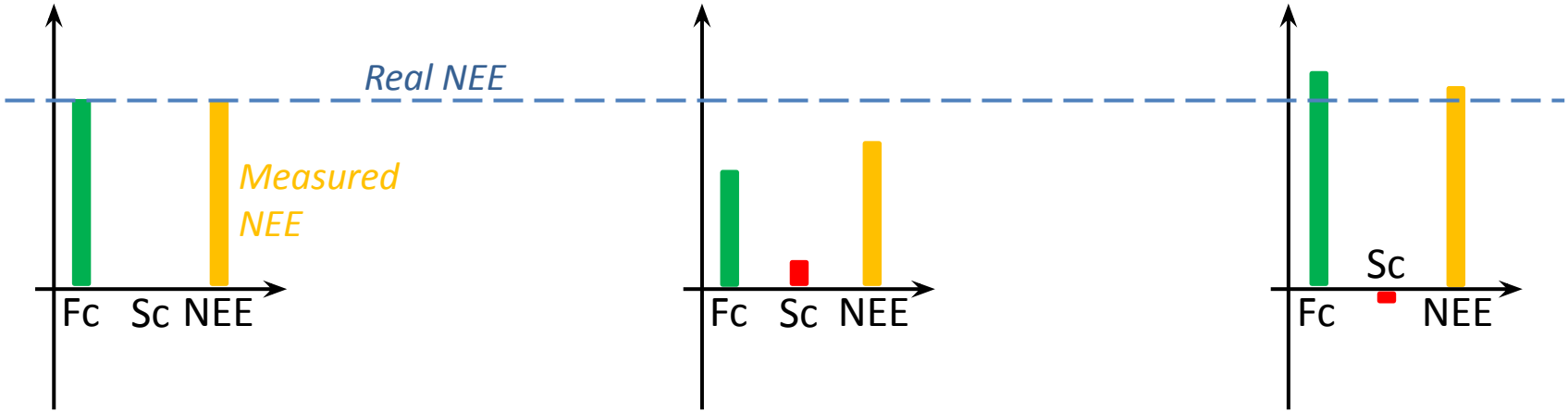
STORAGE + ADVECTION example: nighttime, summer.

I II III

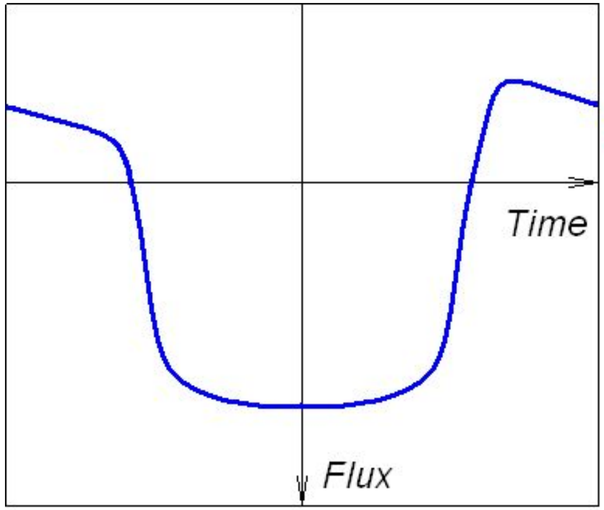


High turbulence (also before) | *Low turbulence* | *High turbulence*

TIME (three half hours)



Storage and advection effects on diurnal pattern



Expected evolution of the biotic flux from an ecosystem with photosintetically active vegetation

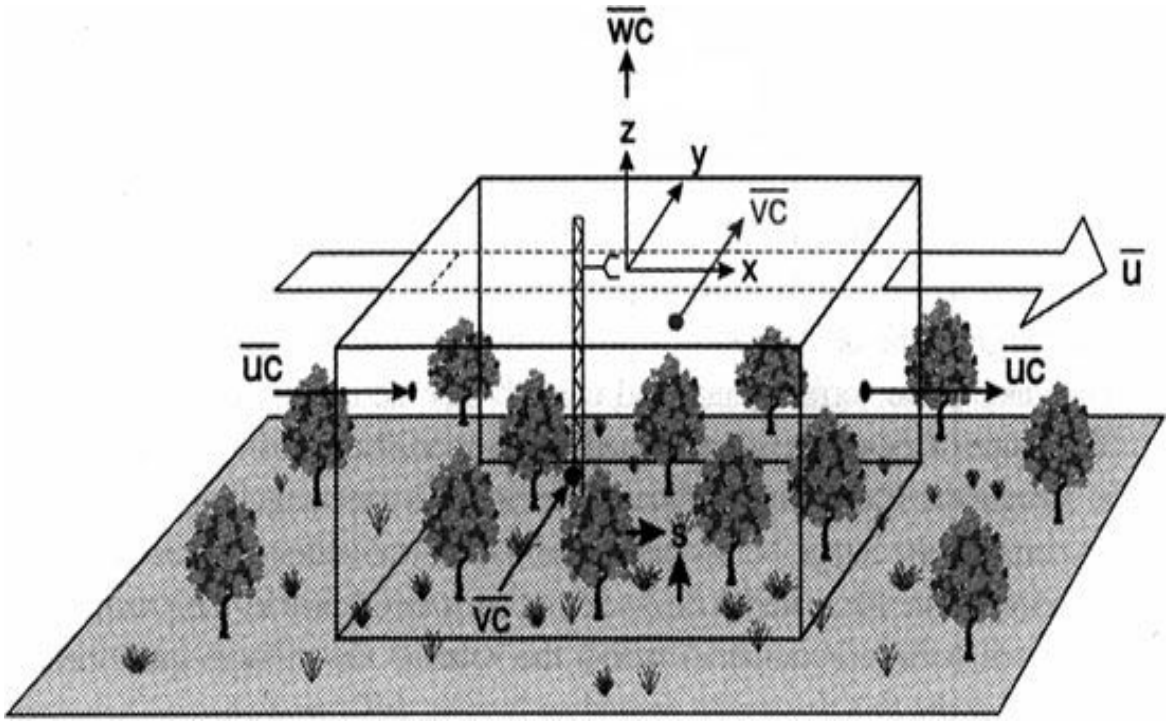
Only storage: during night turbulence decrease, fluxes are underestimated and storage increase. Respiration underestimated, no compensation (at daily timescale!)

advection: if CO2 exported during night, is evacuated by non turbulent transport. Red and green surfaces compensate.

Real situation in most of cases: both storage and non turbulent transport are present (the red and green surfaces don’t compensate).

Storage measurement

We need to measure the CO2 concentration variations inside our reference box (Sc)

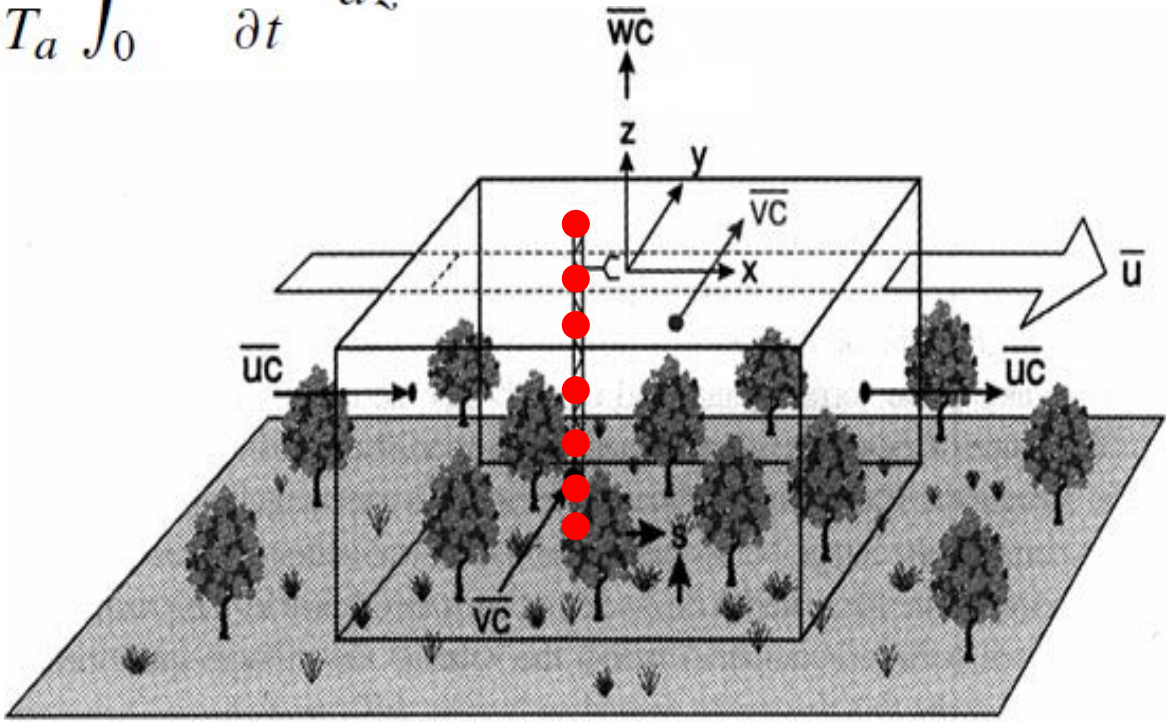


(Finnigan et al. 2003)

Storage measurement

Generally it is calculated using a vertical profile of 5 or more concentration measurement points on towers (logarithmic distribution, denser close to the ground).

$$S_c = \frac{P_a}{R \cdot T_a} \int_0^h \frac{\partial c\,(z)}{\partial t} dz$$



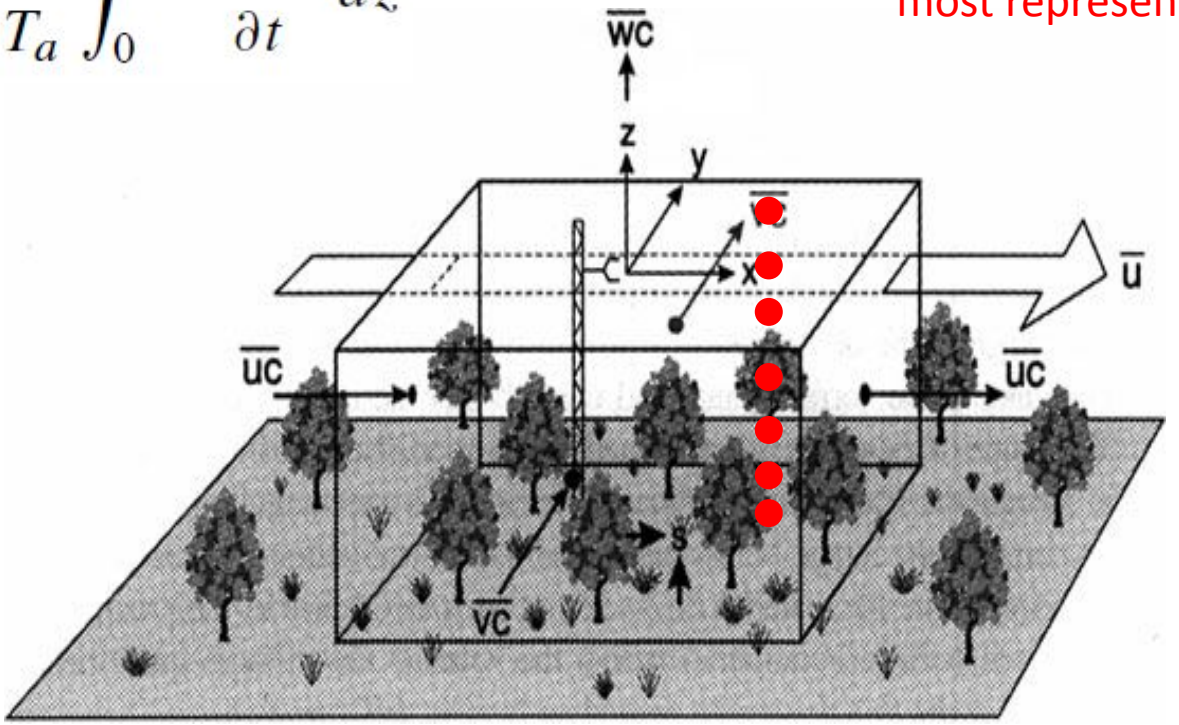
(Finnigan et al. 2003)

Storage measurement

Generally it is calculated using a vertical profile of 5 or more concentration measurement points on towers (logarithmic distribution, denser close to the ground).

$$S_c = \frac{P_a}{R \cdot T_a} \int_0^h \frac{\partial c\,(z)}{\partial t} dz$$

But is the tower position the most representative point?



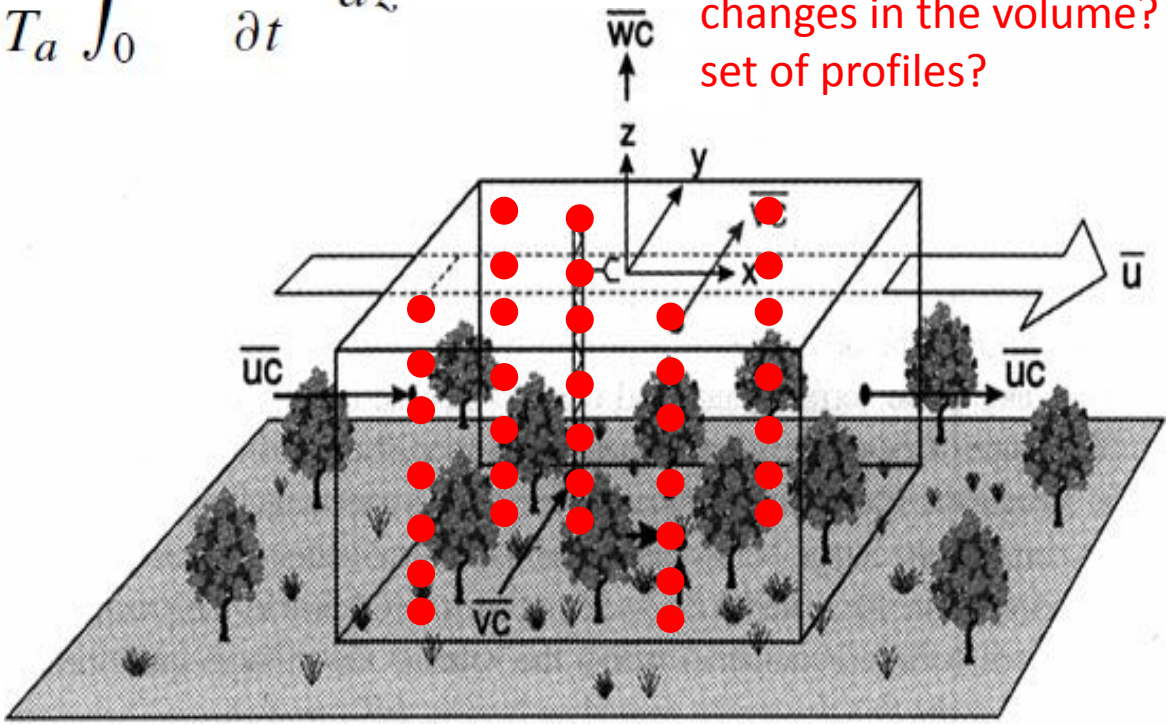
(Finnigan et al. 2003)

Storage measurement

Generally it is calculated using a vertical profile of 5 or more concentration measurement points on towers (logarithmic distribution, denser close to the ground).

$$S_c = \frac{P_a}{R \cdot T_a} \int_0^h \frac{\partial c\,(z)}{\partial t} dz$$

And does it exist a “representative point” where a vertical profile explain changes in the volume? Do we need a set of profiles?



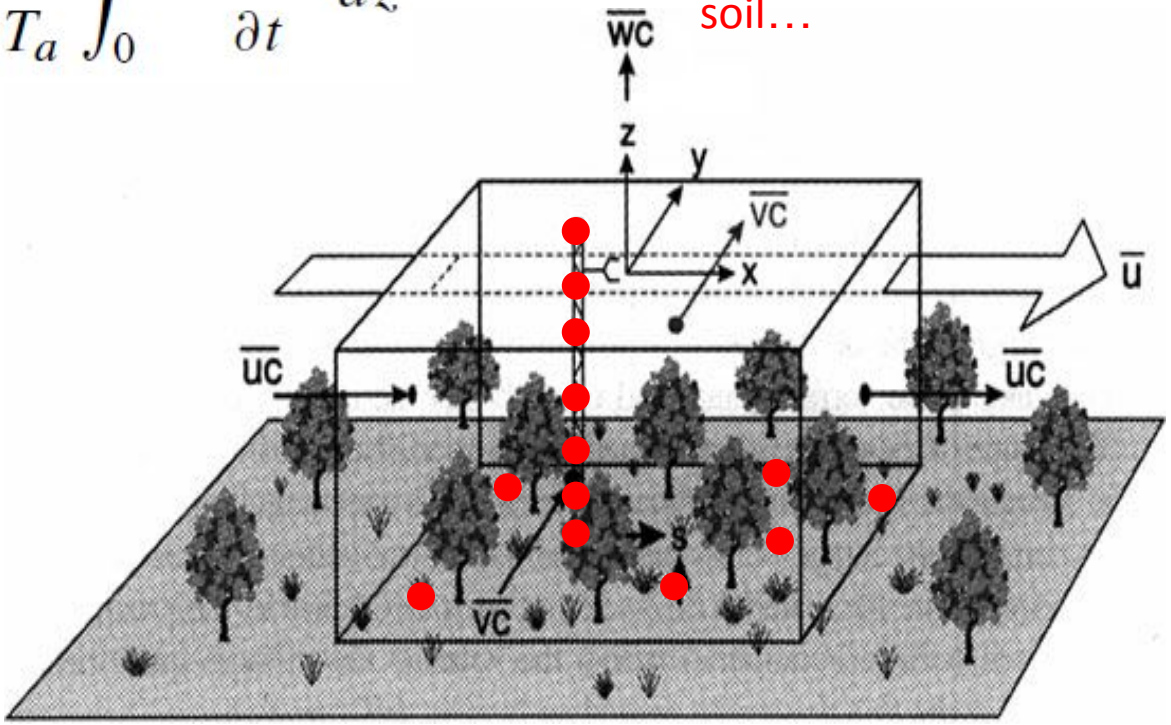
(Finnigan et al. 2003)

Storage measurement

Generally it is calculated using a vertical profile of 5 or more concentration measurement points on towers (logarithmic distribution, denser close to the ground).

$$S_c = \frac{P_a}{R \cdot T_a} \int_0^h \frac{\partial c\,(z)}{\partial t} dz$$

Or may be a compromise solution with denser spatial sampling close to the soil…



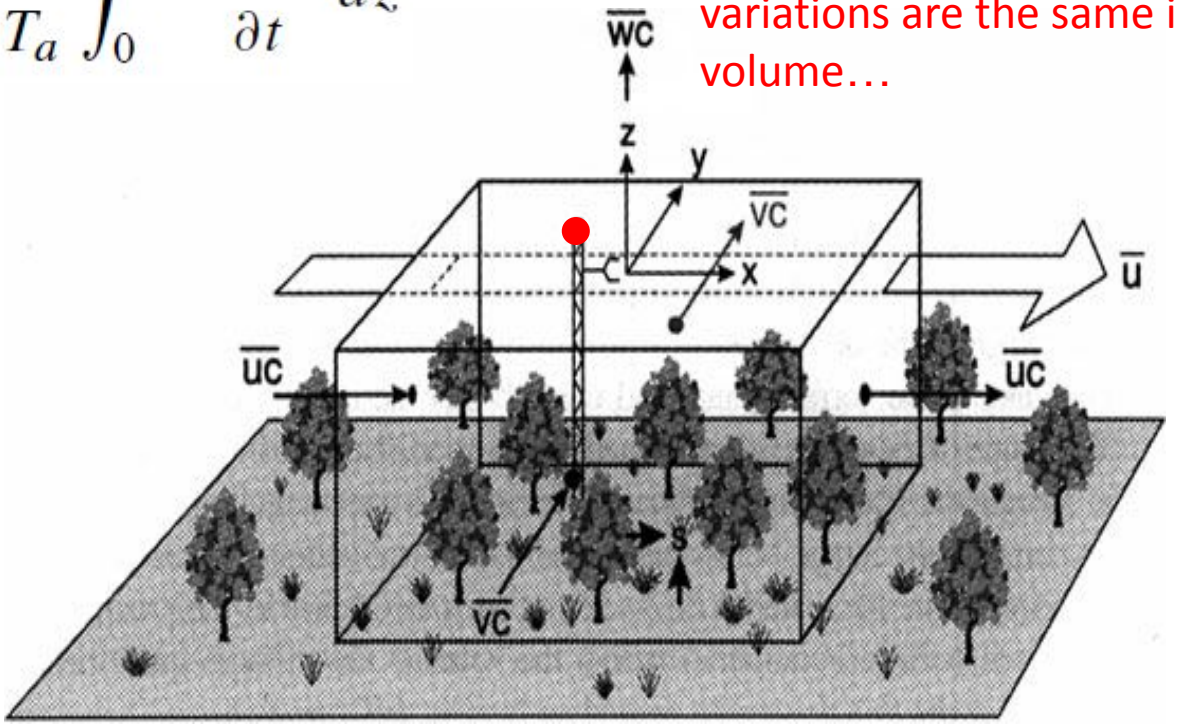
(Finnigan et al. 2003)

Storage measurement

Generally it is calculated using a vertical profile of 5 or more concentration measurement points on towers (logarithmic distribution, denser close to the ground).

$$S_c = \frac{P_a}{R \cdot T_a} \int_0^h \frac{\partial c\,(z)}{\partial t} dz$$

Or very easily just one point and we assume that the concentration variations are the same in the volume…

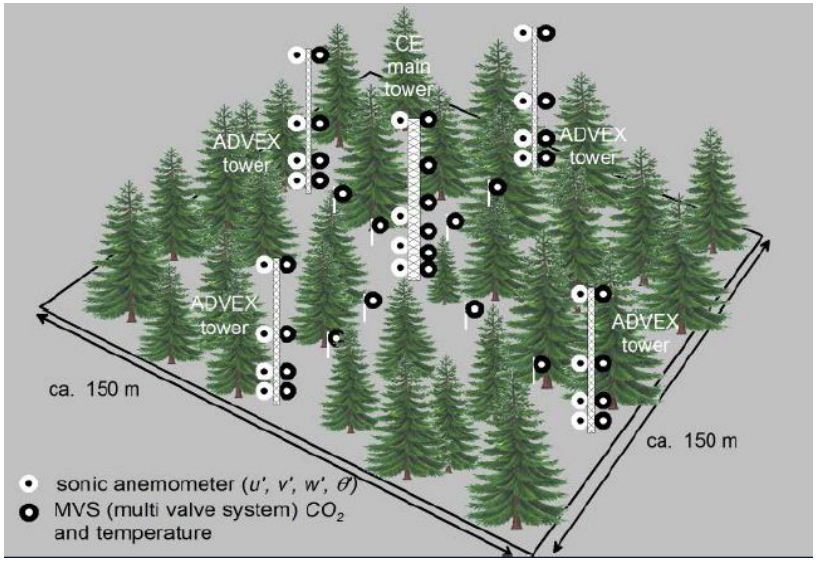


(Finnigan et al. 2003)

Storage measurement – analysis

ADVEX dataset

May-June 2006

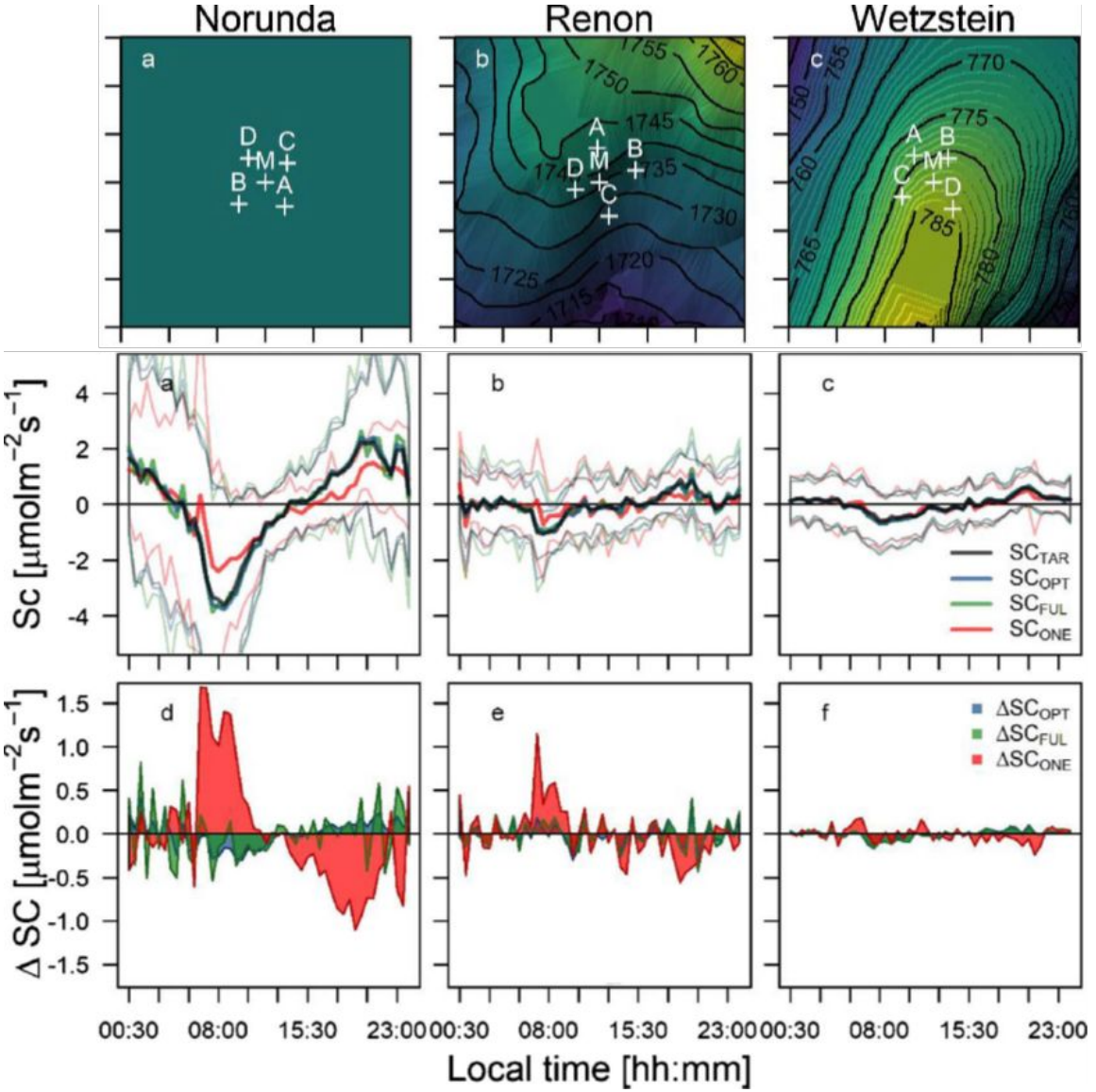


Three sites with multiple towers (to measure advection – see later) all with vertical and horizontal profiles of CO_2

Feigenwinter et al., 2008; Aubinet et al., 2010; Montagnani et al., 2010

Storage flux measurement strategy

How much is it important to correctly measure the storage flux? And which is the best setup compromise?



= *5 profiles for 5 levels*

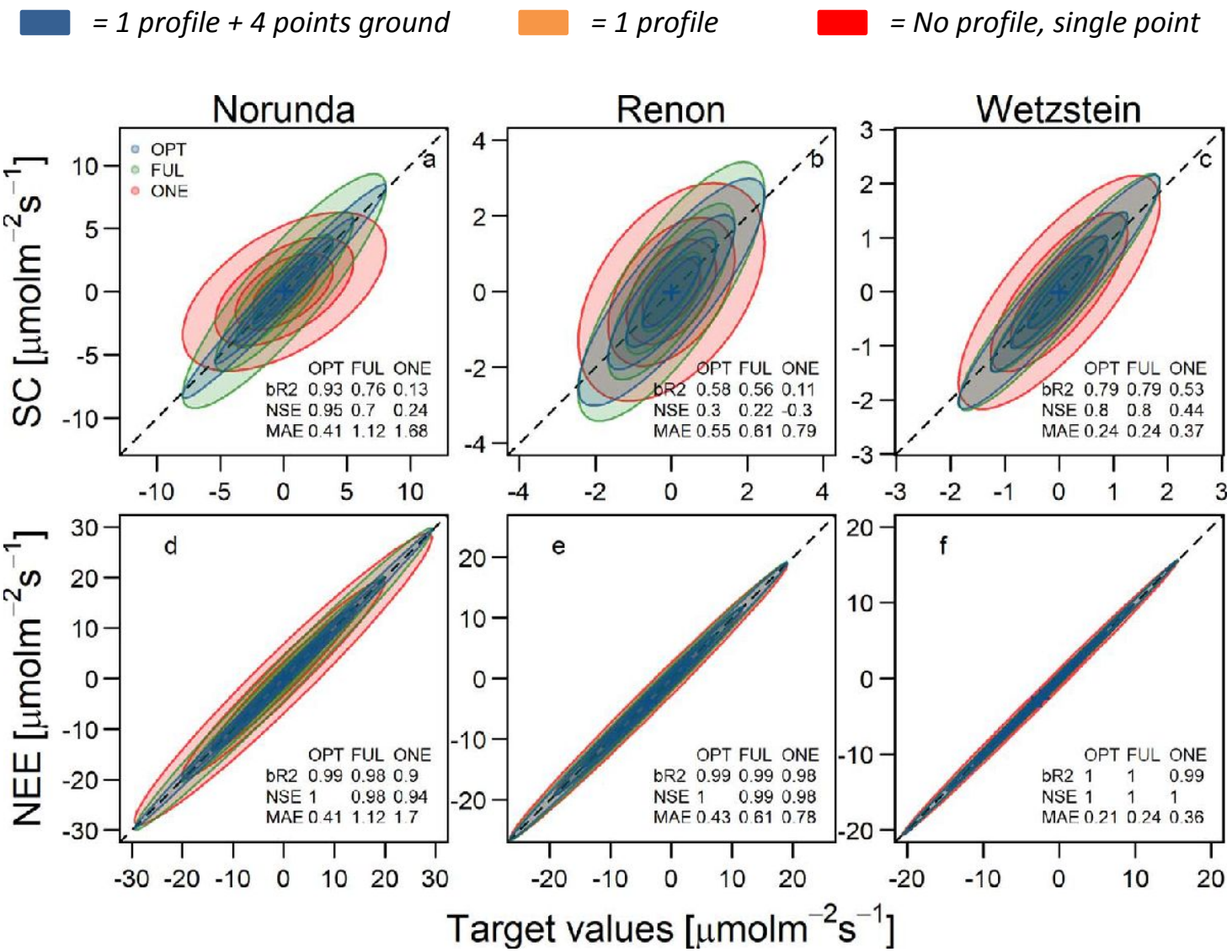
= *1 profile + 4 points ground*

= *1 profile*

= *No profile, single point*

Nicolini et al. 2018

Storage flux measurement strategy

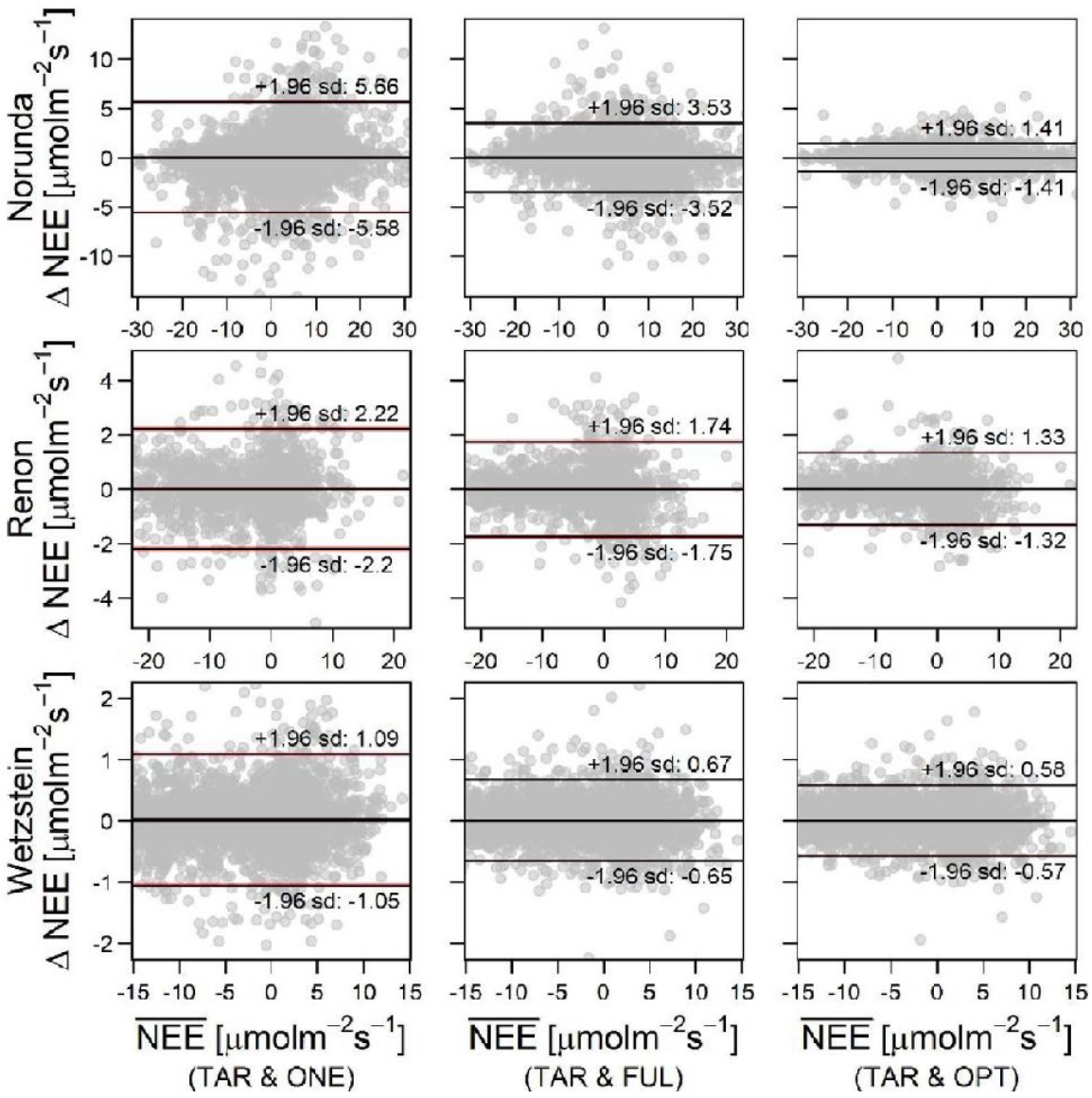


Nicolini et al. 2018

Storage flux measurement strategy

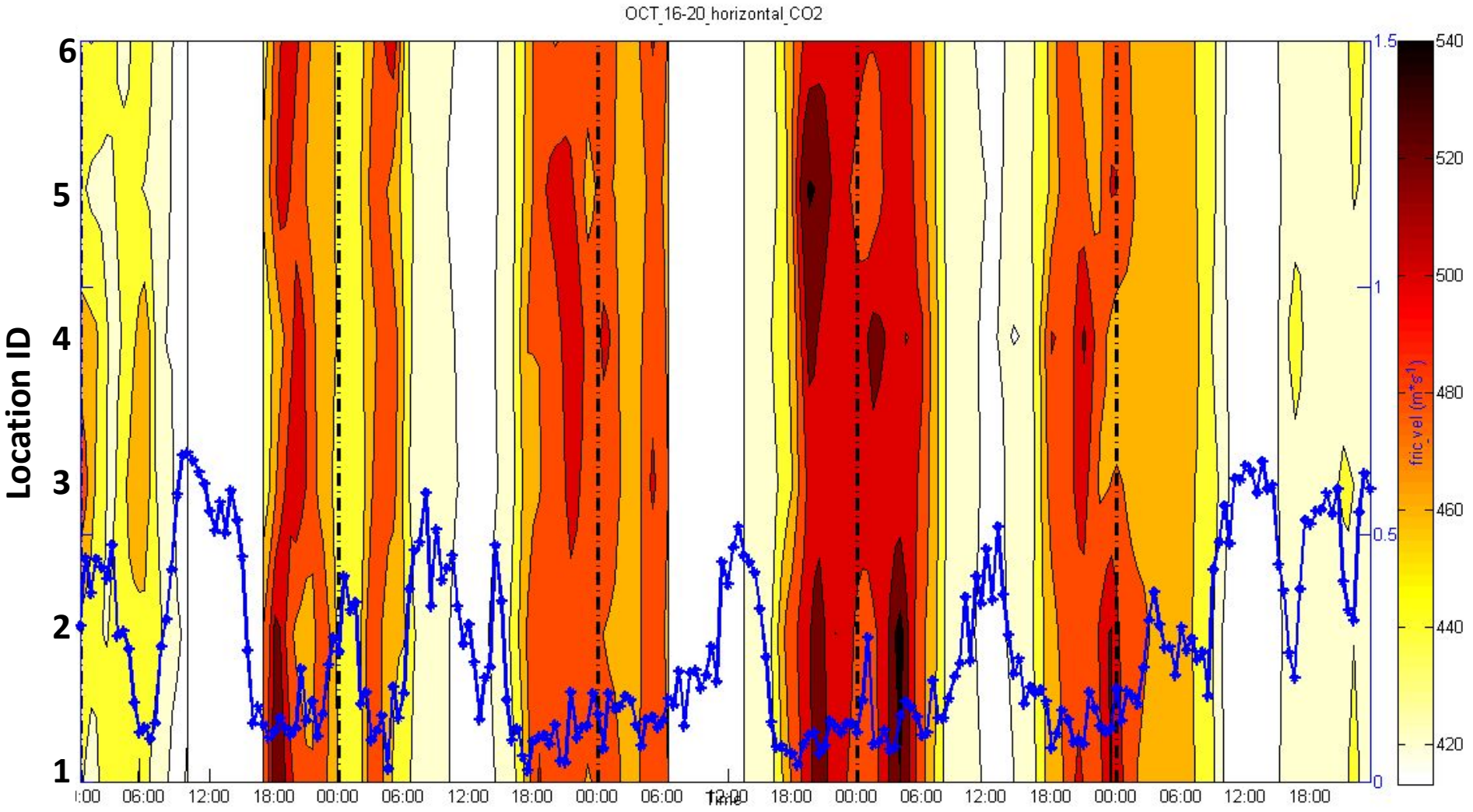
Not as bad as one could expect but a profile is needed…

Are there situations where it can be avodided?



Nicolini et al. 2018

Storage magniture and turbulence



Concentration evolution during 5 days in 6 locations along a transect. In blue u*

data: S. Sabbatini and H. van Asperen

Advection measurement…

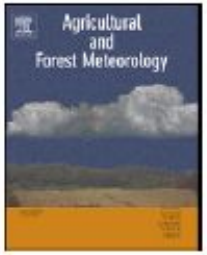
As said the ADVEX data were collected to try to measure and quantify directly the advection but it was impossible due to large scatter (random error).

Agricultural and Forest Meteorology 150 (2010) 655–664

Contents lists available at ScienceDirect

Agricultural and Forest Meteorology

journal homepage: www.elsevier.com/locate/agrformet



Direct advection measurements do not help to solve the night-time CO_2 closure problem: Evidence from three different forests

M. Aubinet[a,*], C. Feigenwinter[a,b], B. Heinesch[a], C. Bernhofer[c], E. Canepa[d], A. Lindroth[e], L. Montagnani[f,g], C. Rebmann[h], P. Sedlak[i], E. Van Gorsel[j]

Advection and ustar filtering

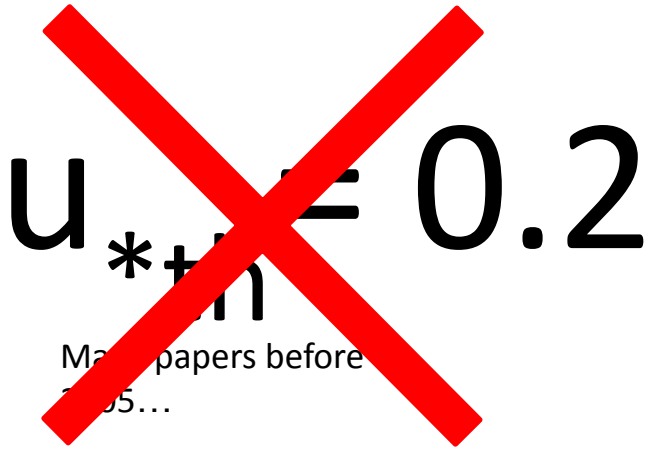
The most used and consolidated method to take into consideration the advection component in the fluxes is the ustar filtering (although it is still controversial)

The general idea is to identify the data that are potentially affected by relevant advection phenomena, remove these data and fill the gaps (if needed) in a later stage.

Ustar is in fact a variable that indicates the turbulence level, so:

larger ustar → more turbulence → more turbulent fluxes → less advective fluxes

We need to identify a threshold of ustar that can be used to define data we have to remove (all the data acquired when ustar < ustar_threshold)



Ustar threshold is site specific, often year specific and must be estimated starting from the data

ustar threshold calculation

General assumptions and idea:

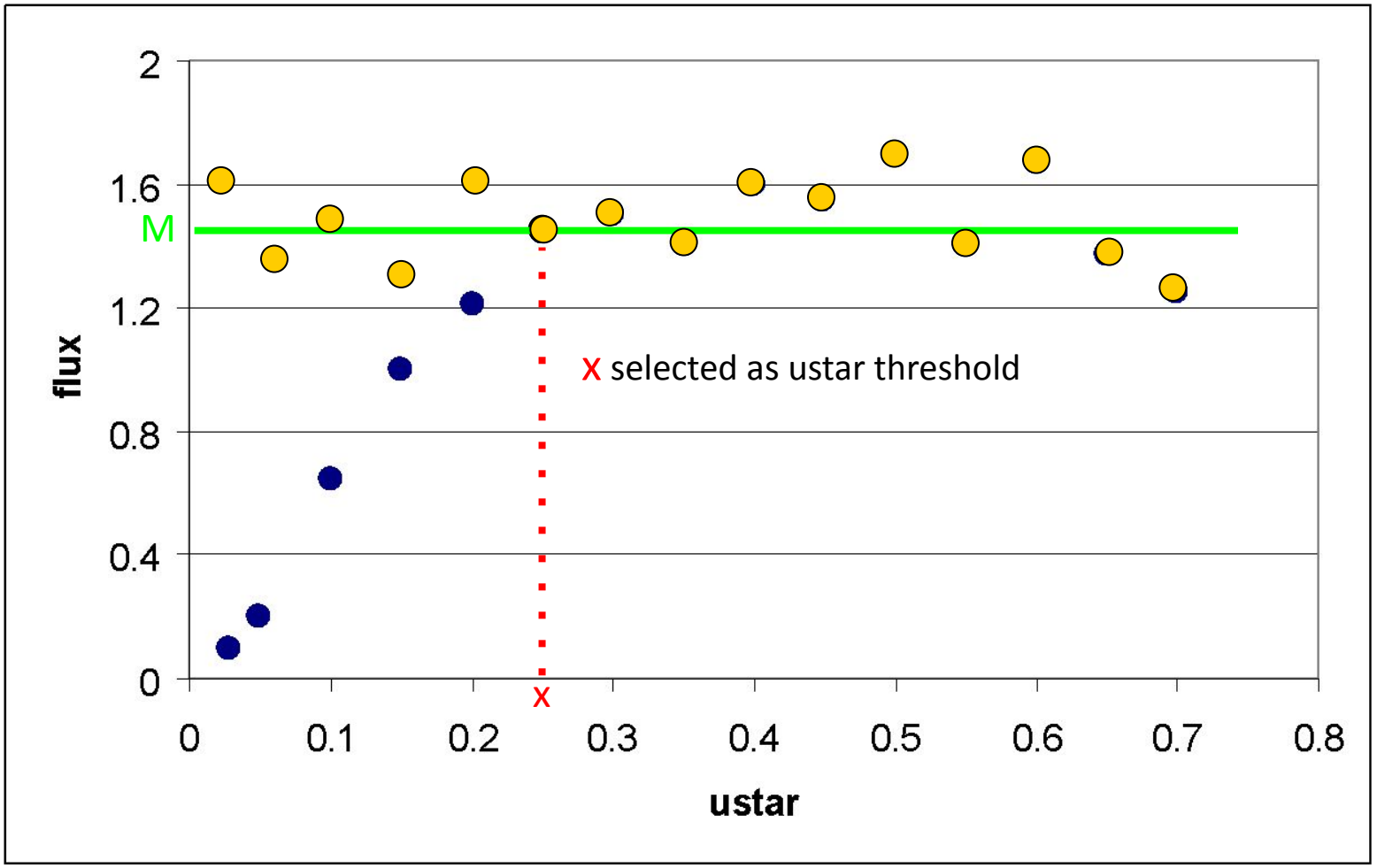
1. During night…
2. …if turbulence is sufficient…
3. …ecosystem respiration is controlled mainly by temperature and time…
4. …so turbulence (ustar) should not affect respiration…
5. … If there is advection, respiration and ustar are not any more independent…
6. …so we can check the respiration-ustar dependency

ustar threshold calculation

General assumptions and idea:

1. During night…
   - Select only nighttime data
2. …if turbulence is sufficient…
   - USTAR
3. …ecosystem respiration is controlled mainly by temperature and time…
   - NEE for similar temperature and similar season
4. …so turbulence (ustar) should not affect respiration…
   - NEE constant respect to USTAR
5. … If there is advection, respiration and ustar are not any more independent…
   - Direct relation USTAR-NEE
6. …so we can check the respiration-ustar dependency
   - Find where (which USTAR) NEE become independent

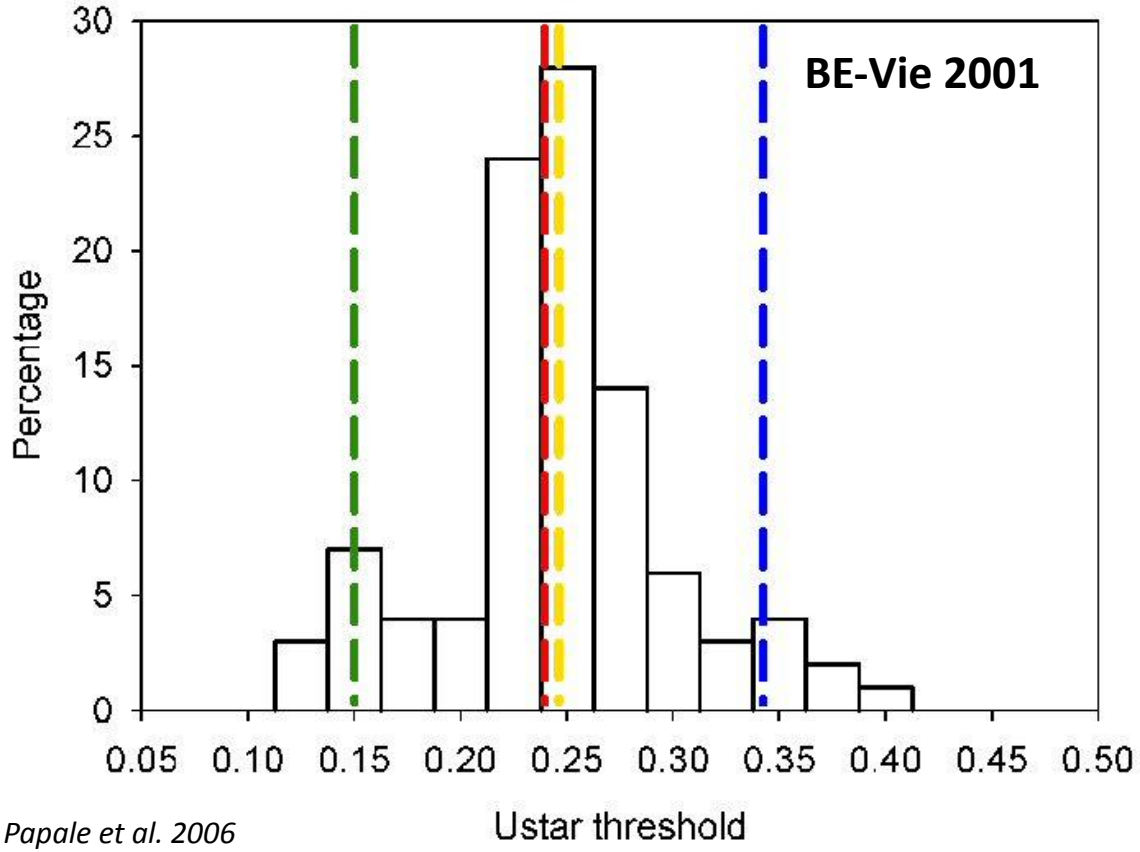
Ustar threshold selection



Can be done manually, however better to use objective, reproducible and automatic methods. Different methods exists (e.g. Reichstein et al. 2005 , Gu et al. 2005, Papale et al. 2007, Barr et al. 2010, Pastorello et al. 2020)

Ustar threshold uncertainty

What is also important is to estimate an uncertainty in the threshold found. Bootstrapping technique is one option that can be used.

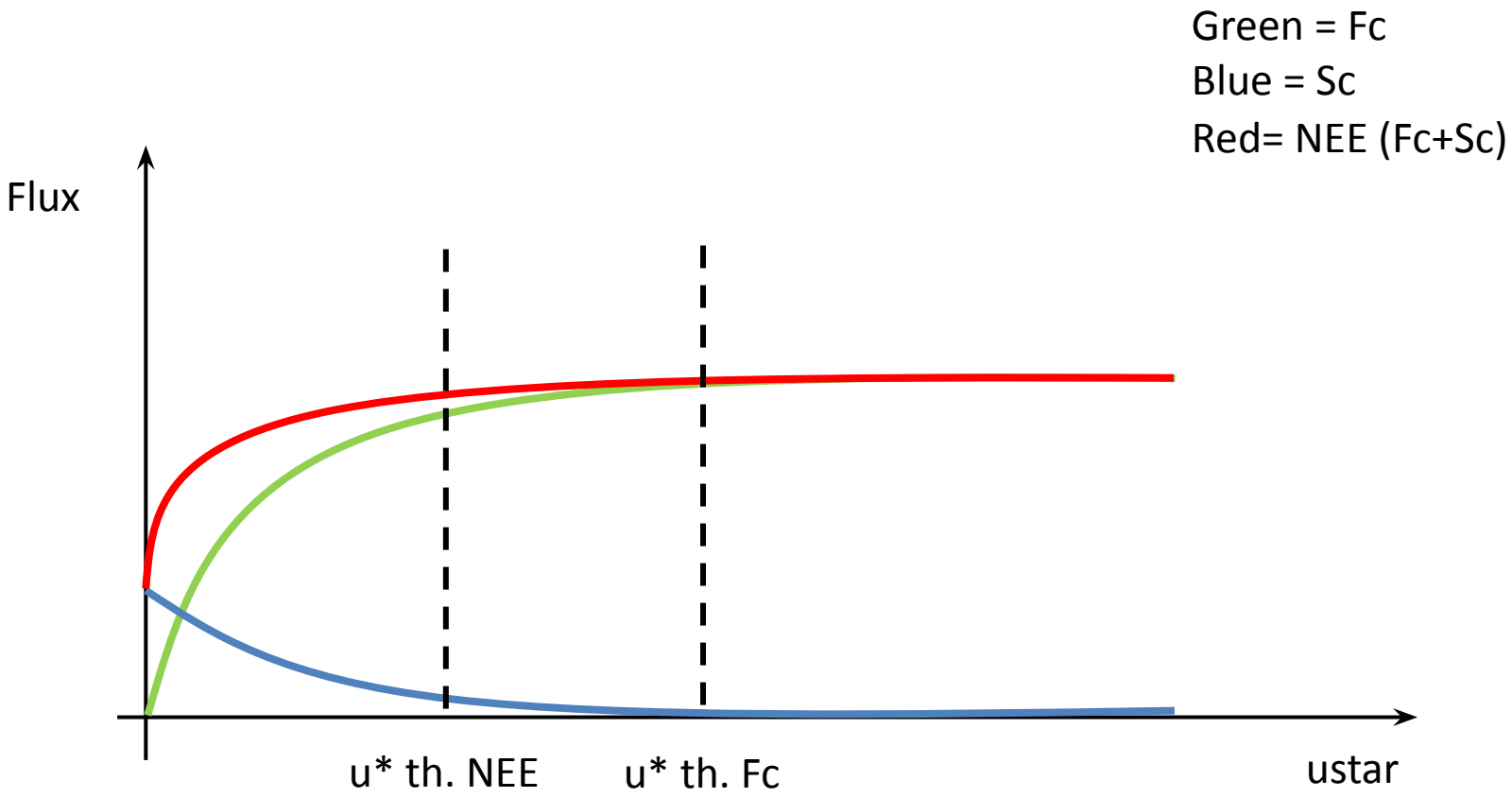


Papale et al. 2006

In the bootstrapping you repeat the analysis several times using different bootstrapped dataset: in each bootstrapping step, the whole dataset is sampled N times (where N = length dataset) where each half-hour can be drawn several times.

5%, Median, 95% percentiles are selected as u*-thresholds to assess the uncertainties

Storage and ustar threshold



So, first the storage correction, then the ustar threshold calculation!

Gap-filling of fluxes timeseries

Just to be clear: gap-filling means imputation, estimation of a missing value in order to obtain a gap-free series of data.

Often the word “gap-filling” is used to indicate the whole post-processing of data: this is not correct.

Gap-filling of fluxes time series

Ustar filtering removes data, some times also a large amount of data. These gaps are added to other missing data periods caused by different reasons

Which gapfilling methods are available? Do I need something complex?

Where are the gaps coming from?

- Power problems
- Instrument problems
- Calibrations
- Quality tests and filtering

NOT RANDOM DISTRIBUTED

For this reason we can not just calculate the average of the integration period

We need other methods

Do we need to fill the gaps?

Gapfilling is not always necessary, but it is necessary when we need to integrate to daily-annual scales

APPLICATION	GAPFILLING?
Functional relations	NO
Budgets	YES
Models parameterization	YES (if daily) / NO (if half-hourly)
Models validation	YES/NO (output time resolution)
…	…

NLR: Non-Linear regressions

Based on parameterized non-linear equations which express (semi-)empirical relationships between the NEE flux and environmental variables such as temperature and light.

Commonly one equation for GPP and one equation for Reco, parameterized using the data available.

$$f(T) = \rho_1^{\rho_2((1/T_{ref})-(1/T))}$$ Arrehnius

$$f(T) = \varphi_1 e^{\varphi_2/(\varphi_3-T)}$$ Lloyd - Taylor

$$f(T) = \frac{\alpha_1}{1 + e^{\alpha_2(\alpha_3-T)}}$$ Logistic (Chen et al 1999)

$$f(D') = \gamma_1 + \gamma_2 \sin(D') + \gamma_3 \cos(D') + \gamma_4 \sin(2D') + \gamma_5 \sin(2D')$$

Seasonal dependence Second-order Fourier function (Hollinger et al 2004). D'=2π x DoY/366

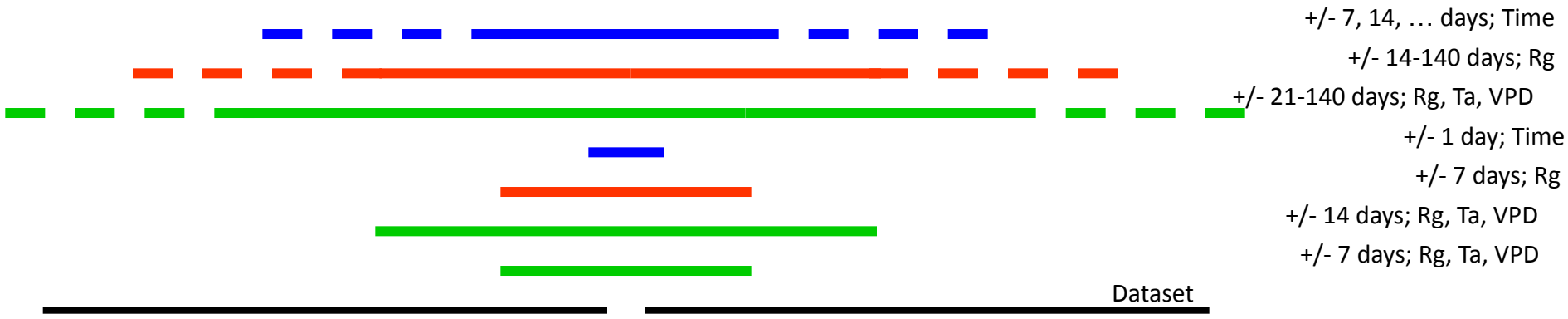
$$\mathrm{GPP} = f(\mathrm{PPFD}) = \frac{\beta_1 \mathrm{PPFD}}{\mathrm{PPFD} + \beta_2}$$ Michaelis - Menten

Regression parameter can be constant for periods varying from one to two months.

LUT/MDS/SPM : Look-up Tables

In a look-up table, the NEE data are binned by variables such as light and temperature presenting similar meteorological conditions, so that a missing NEE value with similar meteorological conditions can be ‘‘looked up’’.

The standard LUT are based on fixed intervals, but there are enhanced methods like Marginal Distribution Sampling (MDS) where the LUT is built around the gap with a dimension and variables that are also not fixed



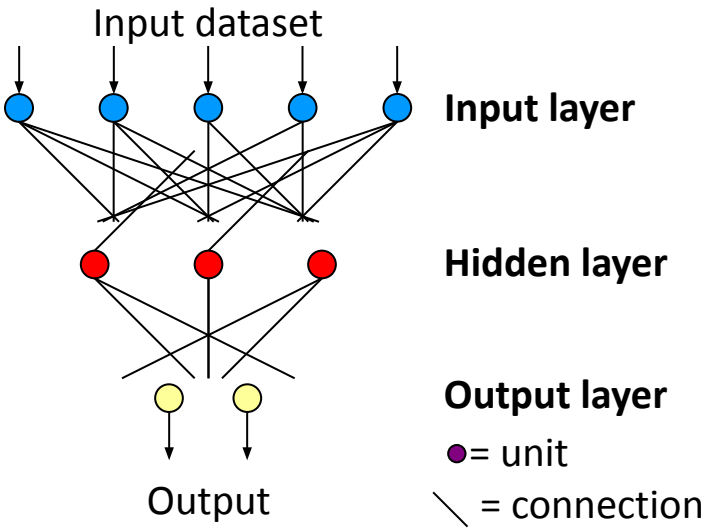
MDV: Mean Diurnal Variation

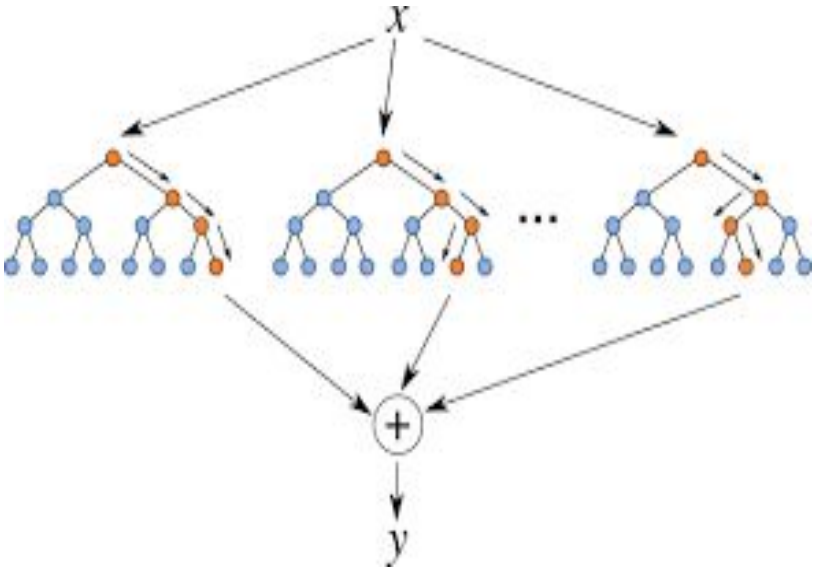
Interpolation technique where the missing NEE value is replaced with the averaged value of the adjacent days at exactly that time of day

Machine learning

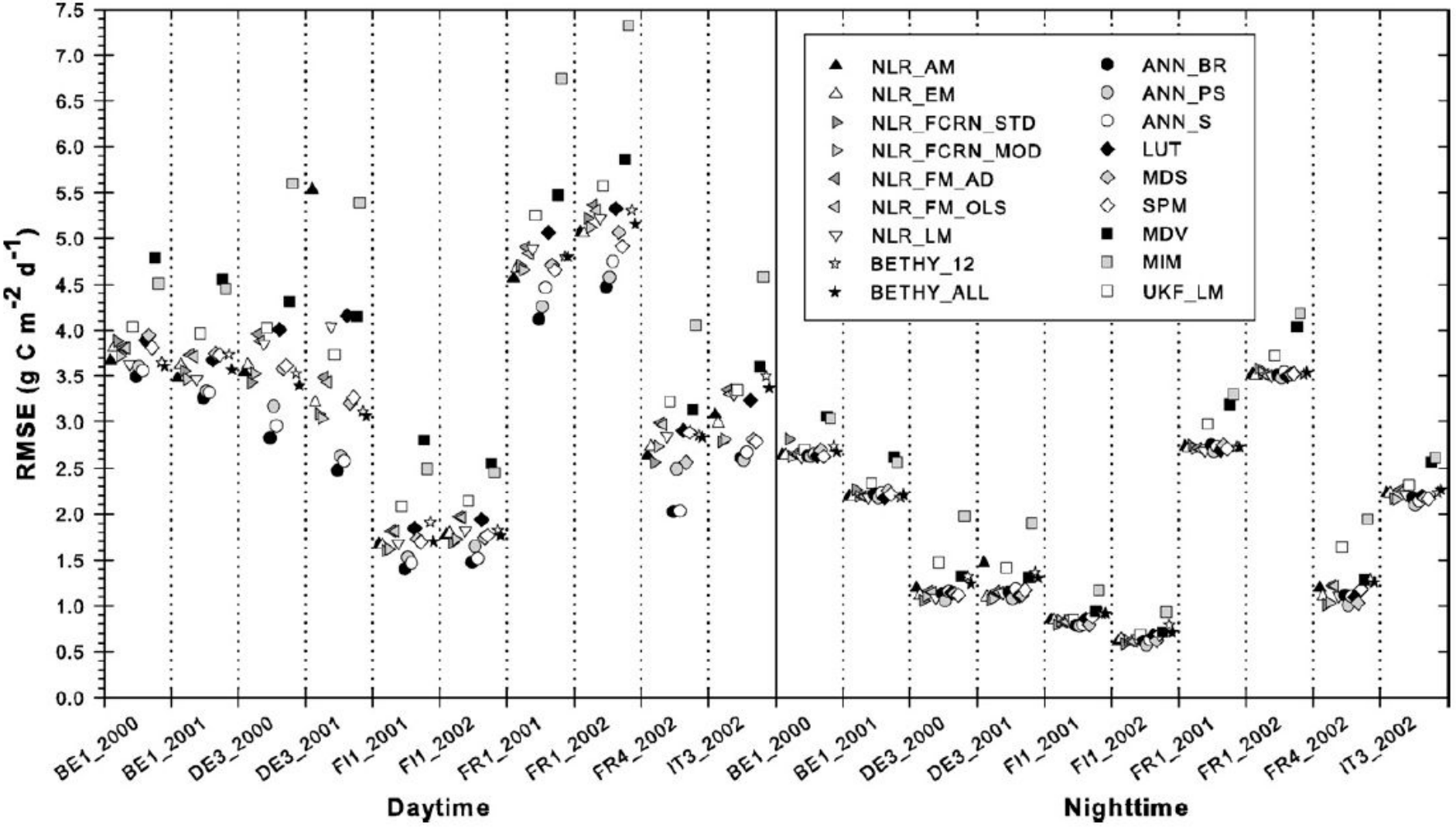
The machine learning are statistical tools, highly flexible and not-linear that can be used to reproduce complex unknown relations between drivers and target (given that the correct drivers are selected)

They are based on training datasets (with drivers and target variables) that are used to parameterize the models. Artificial Neural Networks and Random Forests are two examples of largely used machine learning tools





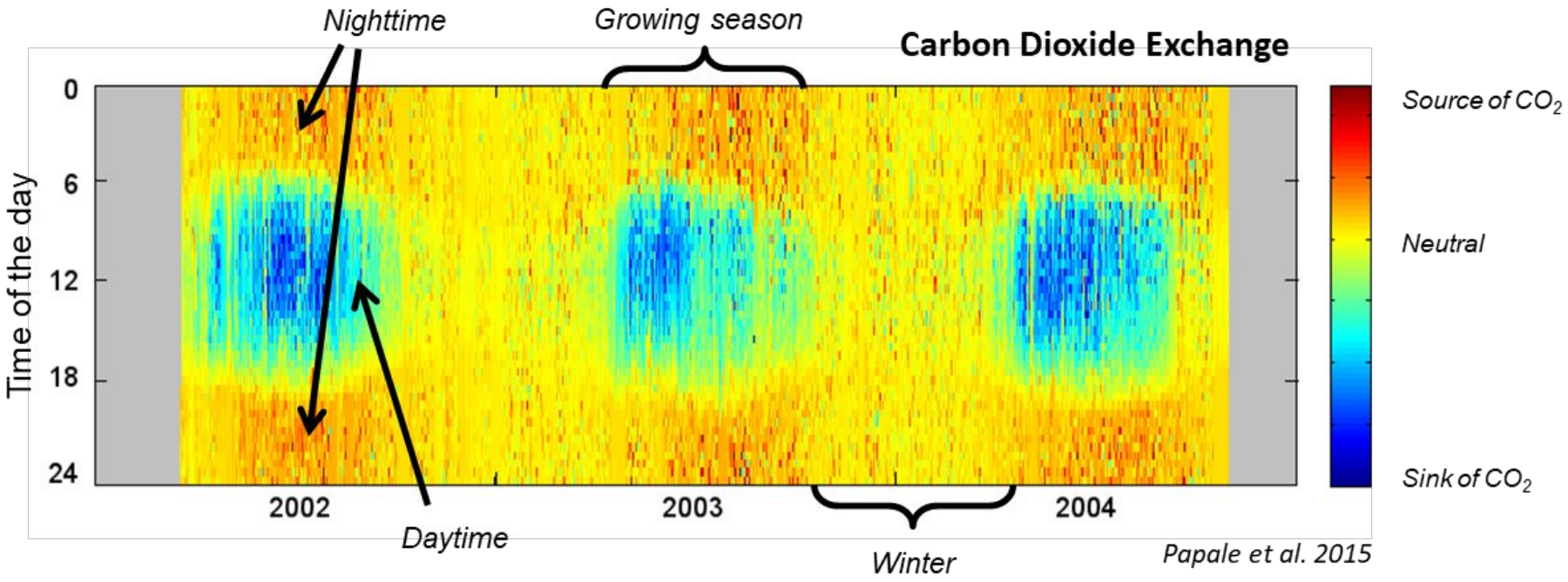
Gapfilling: the gap filling comparison (15 different methods)

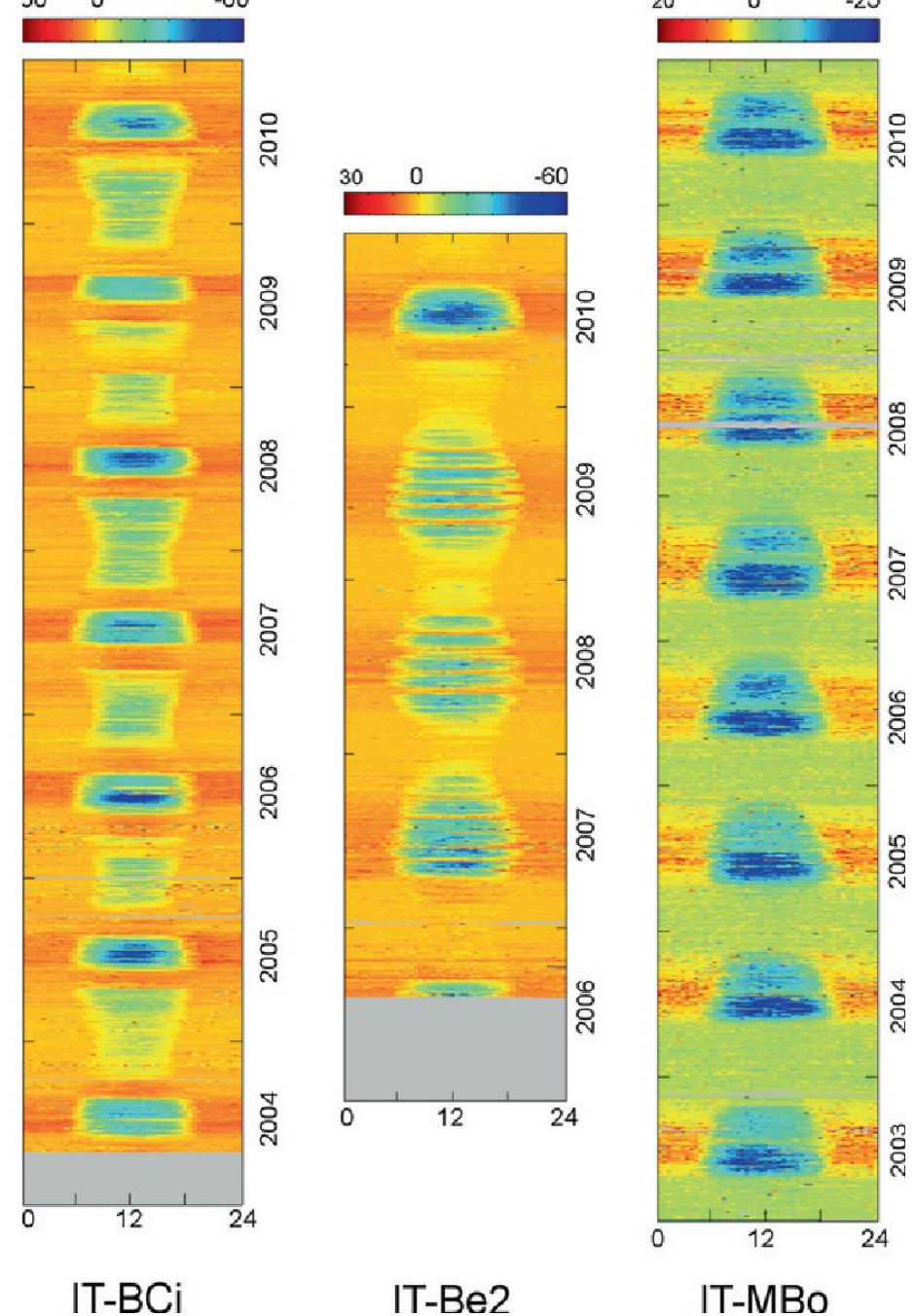


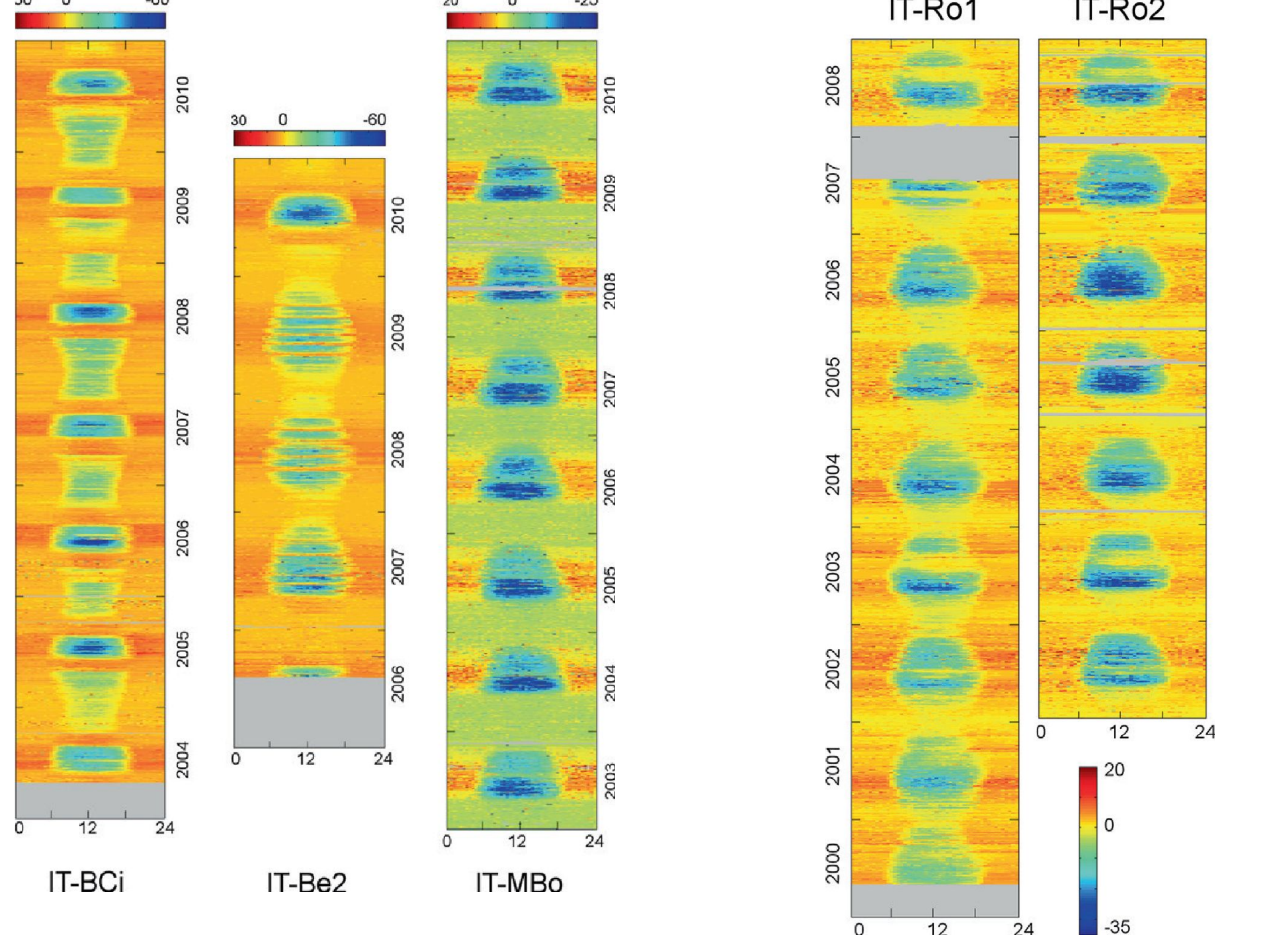
RMSE for different sites and different methods (50 scenarios)

Moffat et al. 2007

The fingerprint plots




30 0 -60
2010
2009
2008
2007
2006
2005
2004
0 12 24
IT-BCi
30 0 -60
2010
2009
2008
2007
2006
0 12 24
IT-Be2
2010
2009
2008
2007
2006
2005
2004
2003
0 12 24
IT-MBo


IT-BCi
IT-Be2
IT-MBo
IT-Ro1
IT-Ro2
30
0
-60
20
0
-25
20
0
-35
0
12
24
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010

Partitioning

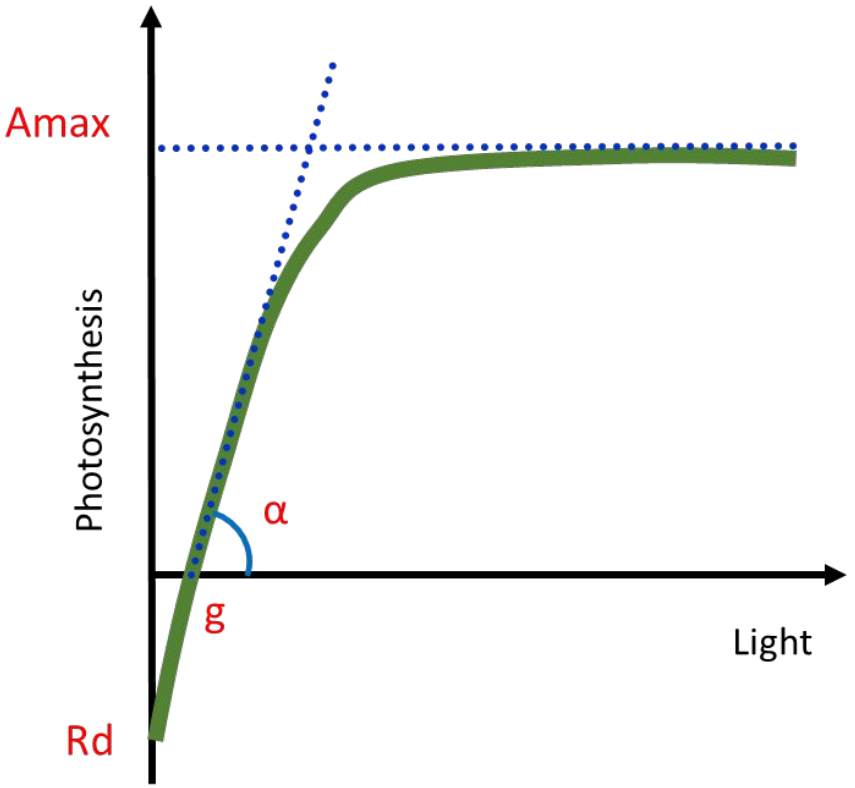
RECO and GPP estimation from eddy data

With eddy covariance we are measuring NEE but using partitioning methods it is possible to assess also the two main components photosynthesis (GPP) and ecosystem respiration (RECO).

There are two main approaches generally used:

- Based on night time data, extrapolating RECO measured at night to daytime (Reichstein et al. 2005)
- From day time data, using a two components model of NEE with light-response curve and exponential function for respiration (Lasslop et al. 2010)

Photosynthesis and light



Amax = Maximum assimilation

α = quantum efficiency

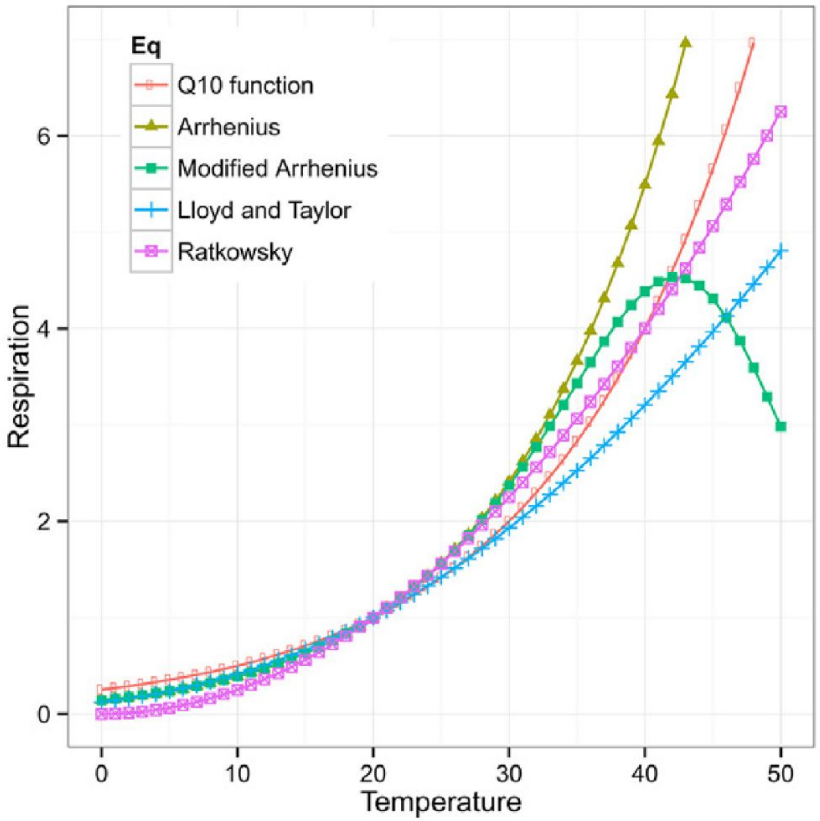
Rd = dark respiration of leaves

g = compensation point

$$Phot = \frac{\alpha * PAR * Amax}{\alpha * PAR + Amax}$$

Michaelis - Menten
1913

Temperature effect on respiration



Different models proposed for the temperature-respiration relation.

Lloyd and Taylor (1994):

$$R_{\mathrm{eco}} = R_{\mathrm{ref}} \mathrm{e}^{E_0(1/(T_{\mathrm{ref}}-T_0)-1/(T-T_0))}$$

R_{eco} = ecosystem respiration
R_{ref} = respiration at T_{ref}
E_0 = activation energy
T_{ref} = reference temperature
T_0 = -42.06 °C

Night-time method

$$R_{eco} = R_{ref} e^{E_0 (1/(T_{ref} - T_0) - 1/(T - T_0))}$$

(Lloyd & Taylor, 1994)

T_0 = -46.02 °C

T_{ref} = 10 °C

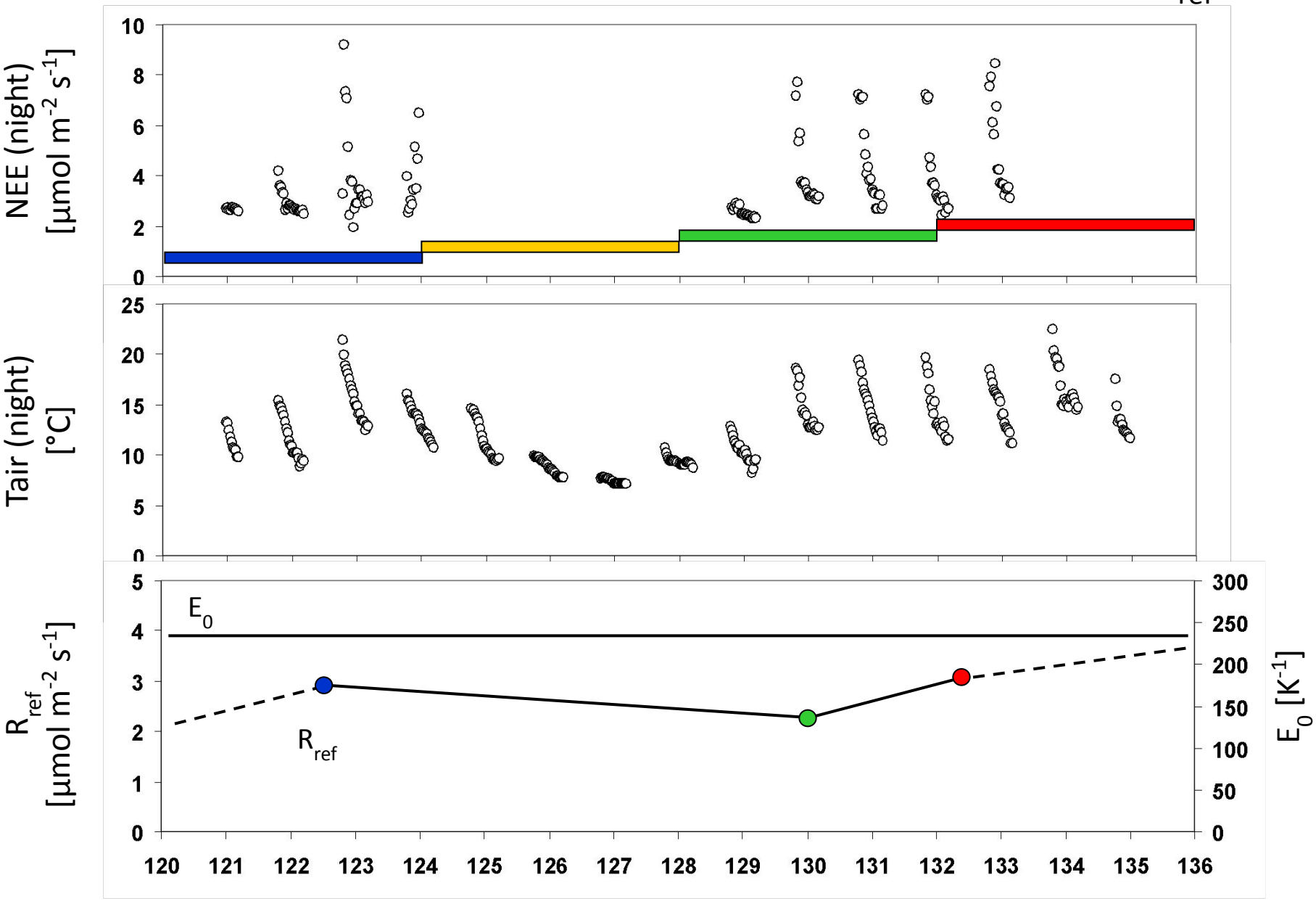
Nighttime eddy covariance measurements

Estimation of E_0 (temperature sensitivity) and R_{ref}

First estimate E_0 then R_{ref} both from short time windows

Night-time method

Estimation of temperature independent respiration level R_{ref}



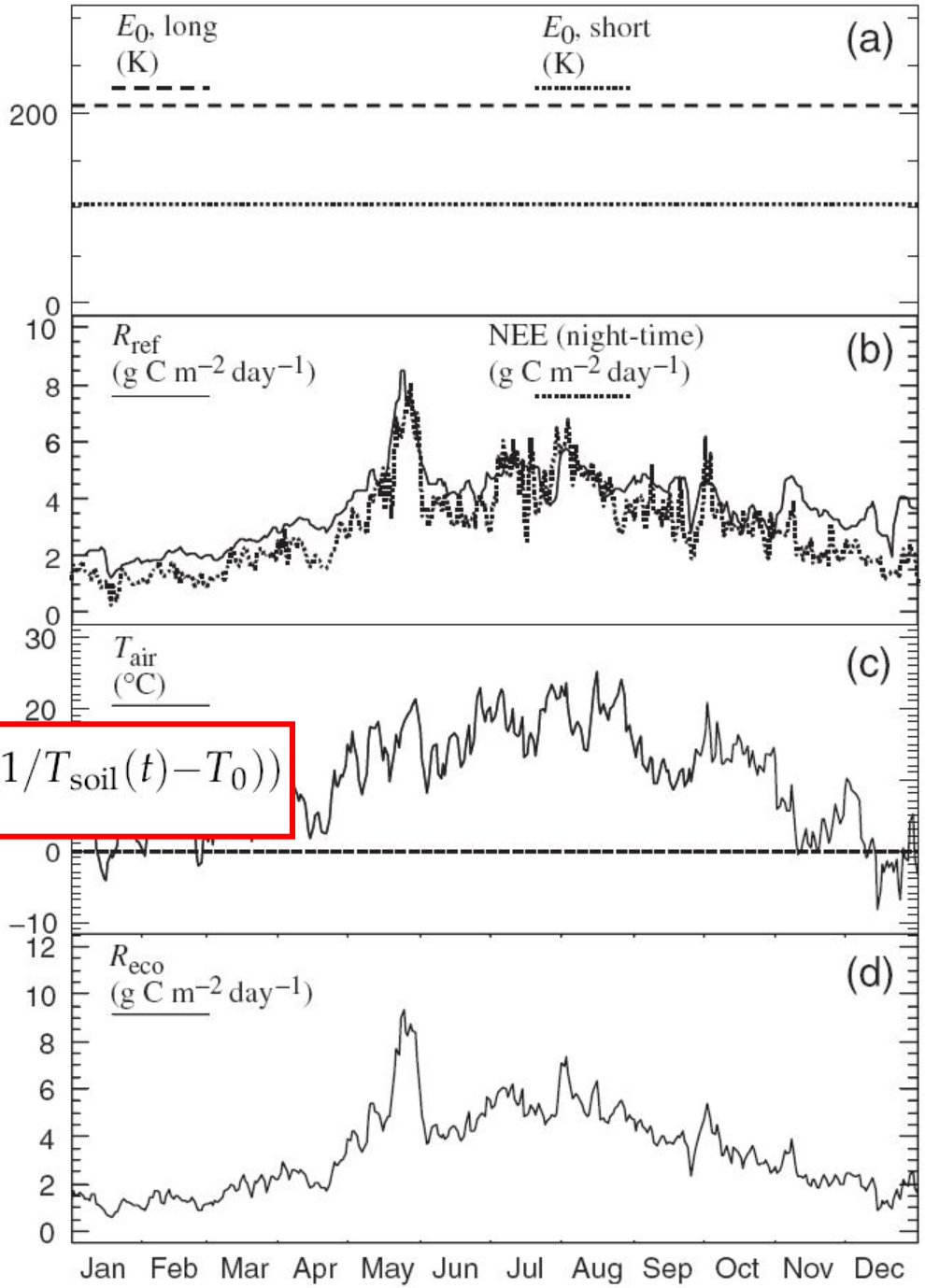
Night-time method

R_{eco} estimate

Seasonal variation
(phenology, SWC, microbial dynamic…)

$$R_{eco}(t) = R_{ref}(t)e^{E_0(1/(T_{ref}-T_0)-(1/T_{soil}(t)-T_0))}$$

t = time dependent



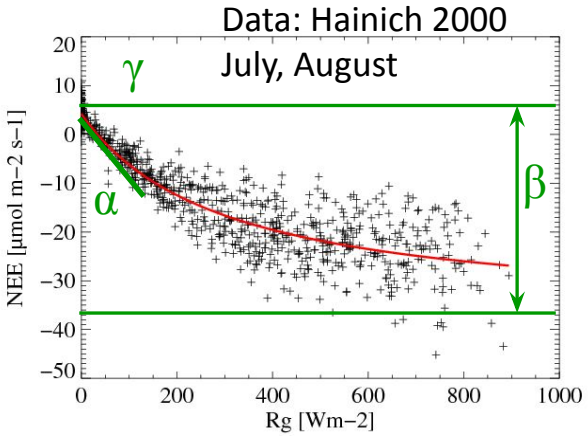
(Reichstein et al., 2005)

Daytime based partitioning algorithm

$$NEE = -\frac{\alpha\beta R_g}{\alpha R_g + \beta} + r_b \exp\left(E_0\left(\frac{1}{T_{ref} - T_0} - \frac{1}{T_{obs} - T_0}\right)\right)$$

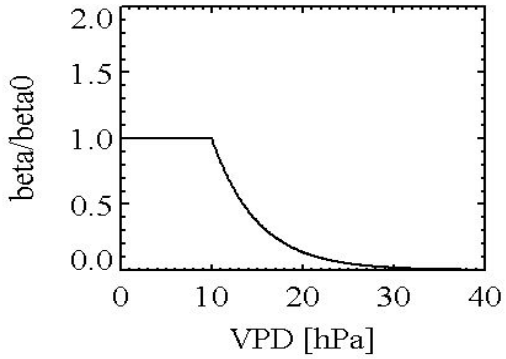
Light response function

Lloyd&Taylor respiration model



Water stress effect on stomata closure (photosynthesis reduction) is added varying the maximum C uptake with VPD (VPD_0 = 10 hPa)

$$\beta = \begin{cases} \beta_0 \cdot \exp(-k\,(VPD - VPD_0)) & \text{if } VPD > VPD_0 \\ \beta_0 & \text{otherwise} \end{cases}$$



There are five parameters to estimate (equifinality problem), E_0 estimated using night-time data, the others using four days mobile windows on daytime data only.

Lasslop et al. GCB (2010) for details

The negative GPP…

When partitioning based on night-time respiration extrapolation is used, GPP is calculated as:

$$GPP = Reco - NEE$$

Since during night there is only respiration

$$NEE_{night} = Reco \quad \textit{and} \quad GPP_{night} = 0$$

However it happens that

1) $NEE_{night} > Reco$ and from the calculation we have **GPP during night**
2) $Reco - NEE_{day} < 0$ and we have **negative GPP**

What to do?

Consider the random uncertainty and don't filter the data otherwise a bias is introduced!!

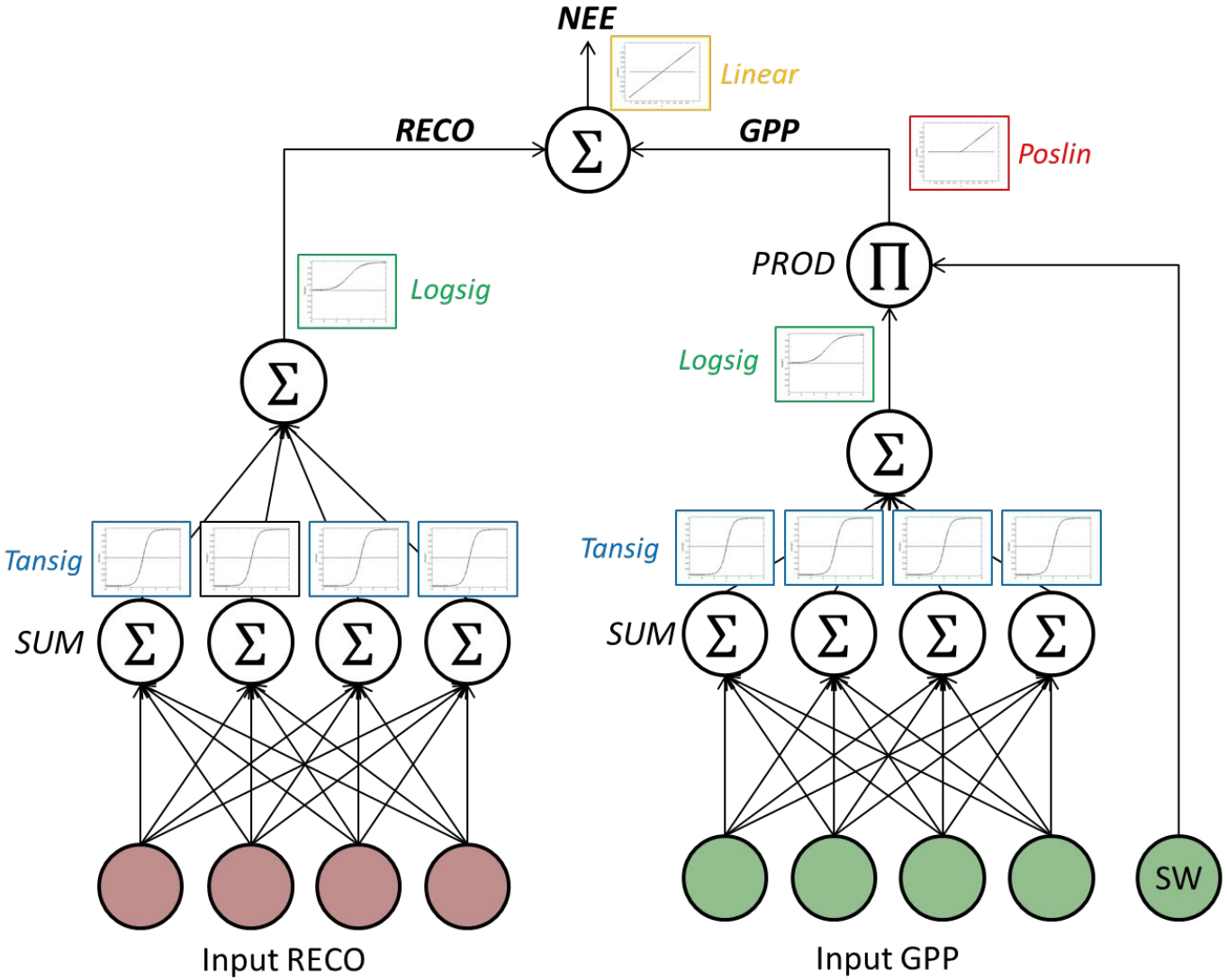
Partitioning using machine learning methods

RECO drivers:

- TA, TS, SWC, WS, WD
- DOY
- Daily NEE_{night} mean

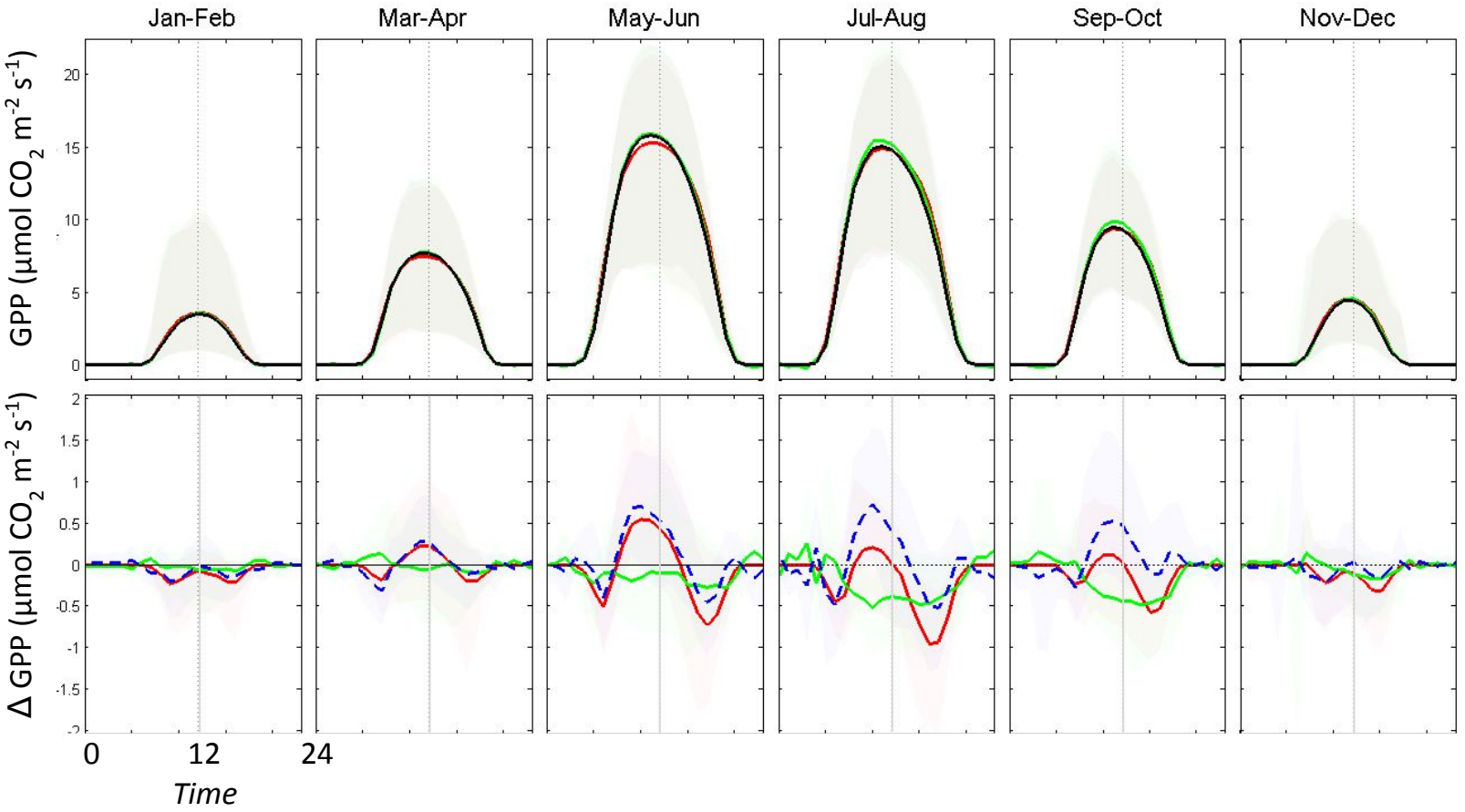
GPP drivers:

- TA, SWC, WS, WD
- Daily NEE_{day} mean
- SW_POT (transformed)
- + SW_IN



Tramontana et al. 2020

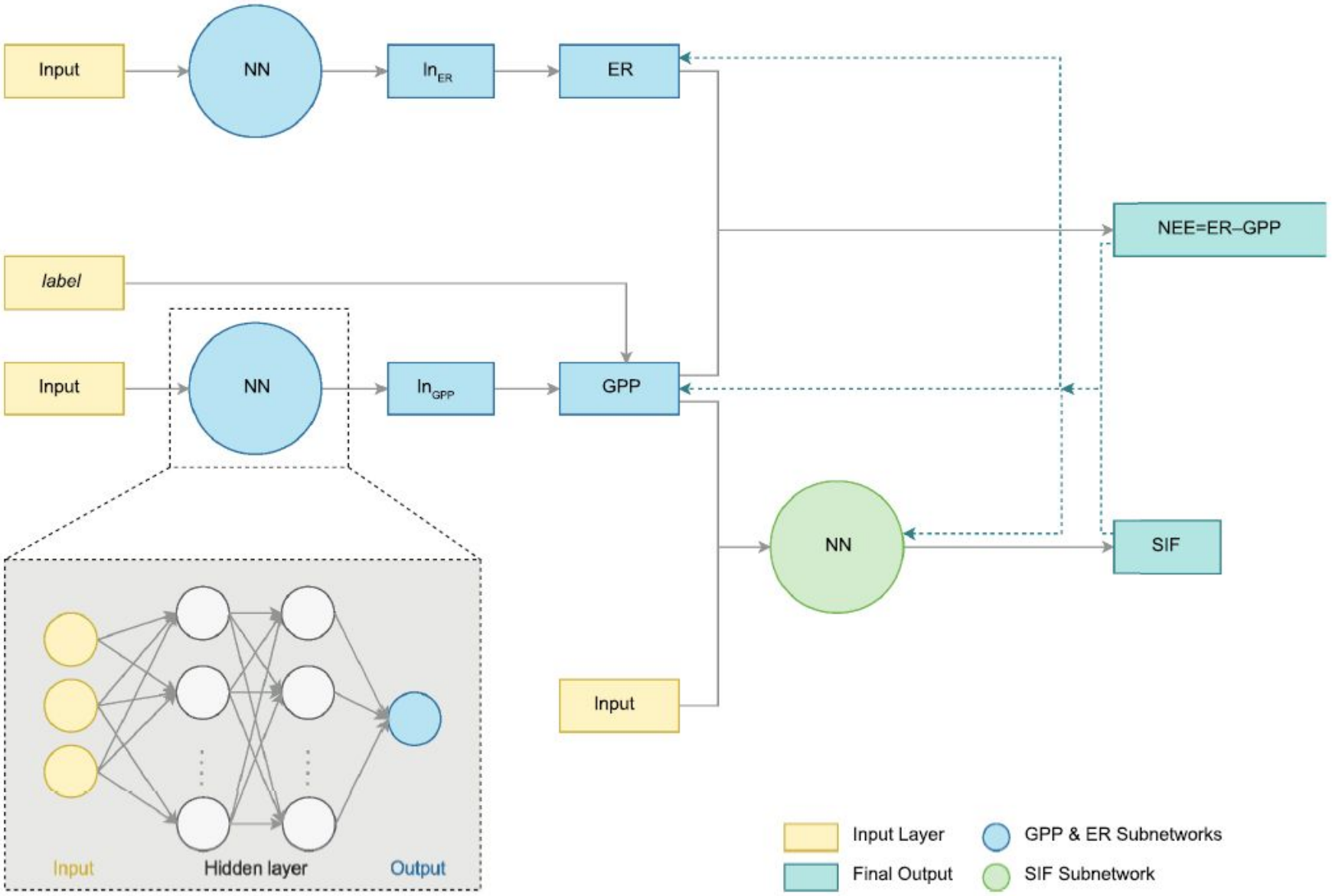
Results: mean diurnal cycle agreement GPP



Higher fluxes predicted by NT and also NN_{cust} in the central part of the day respect to DT method opposite in the afternoon (in particular in NN_{cust}) – VPD effect? Prescribed response for GPP in DT

Tramontana et al. 2020

Machine learning and SIF additional constrain

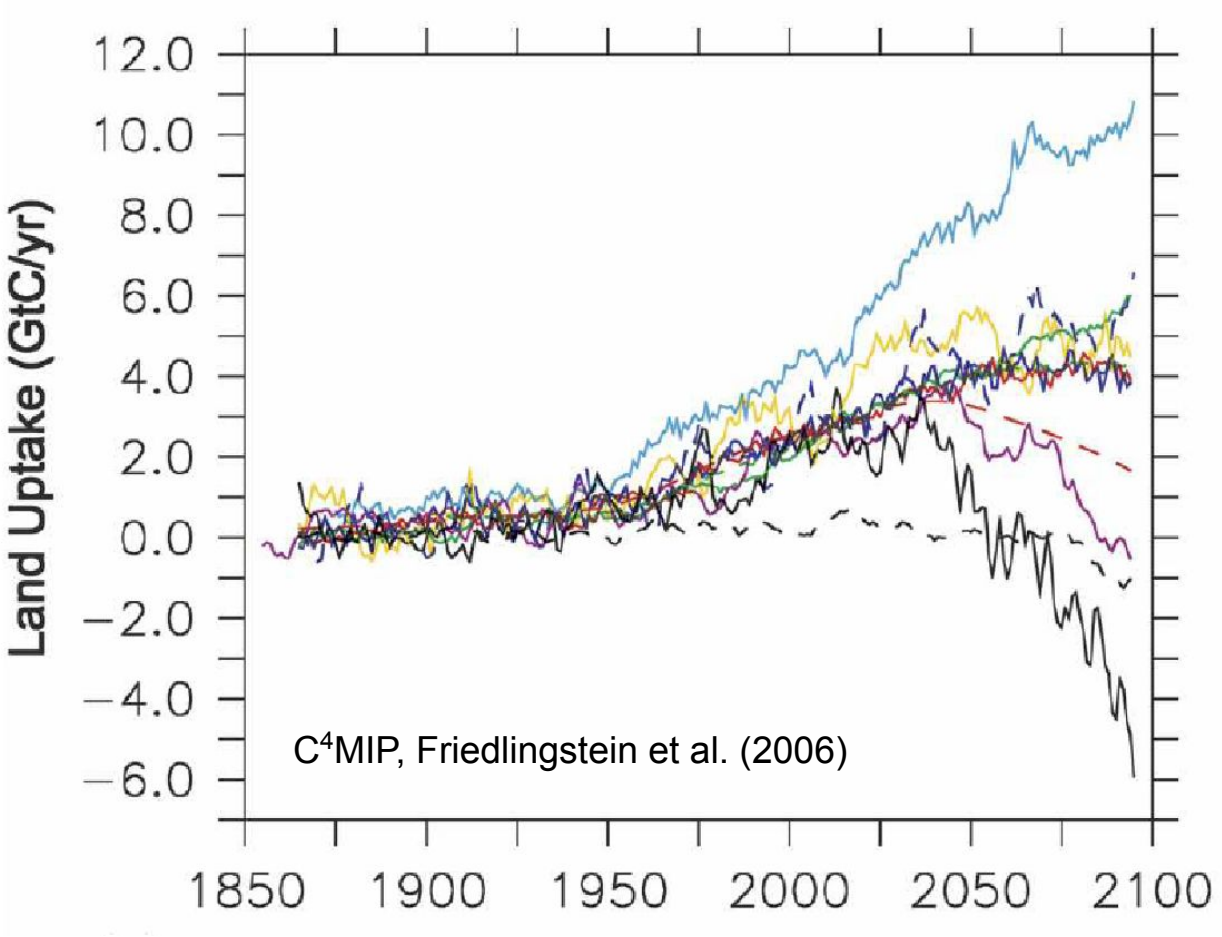


Zhan et al. 2022

Gap-filling and partitioning

What happens in case of management, disturbances, heterogeneity?

Remember: partitioning is a modeling exercise and perfect models don't exist. The uncertainty in the partitioning is in line with other models and important to be considered…

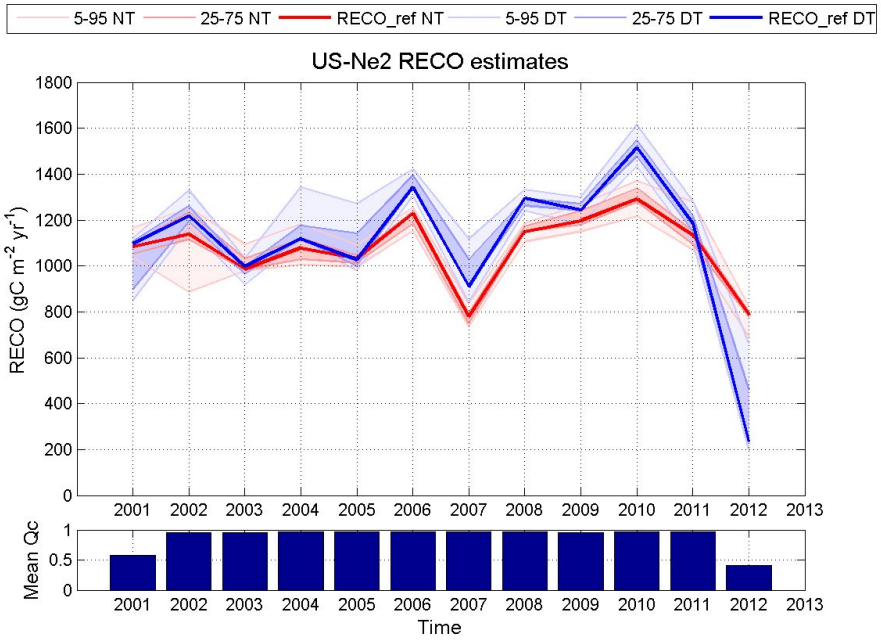


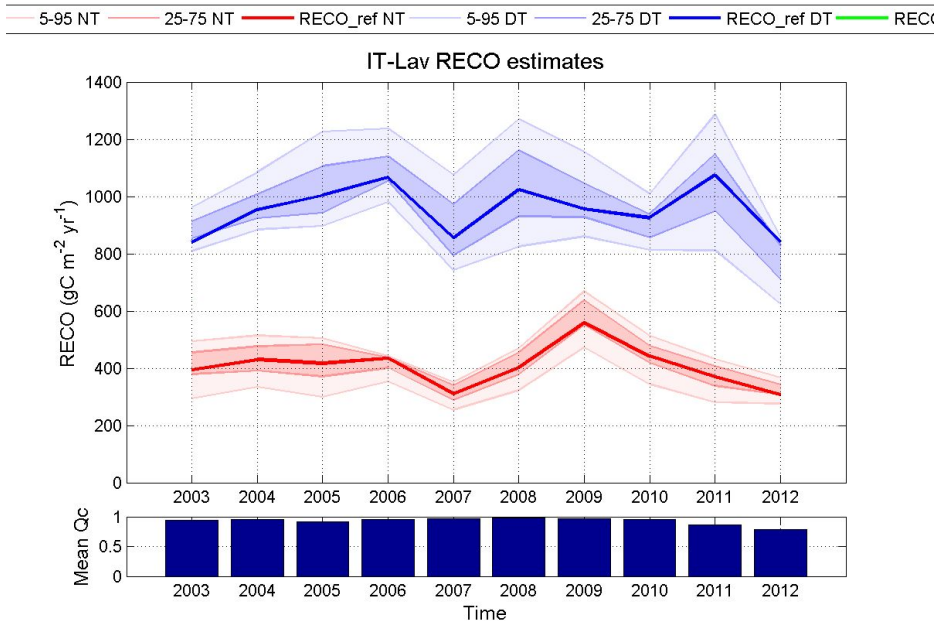
Where is the uncertainty?

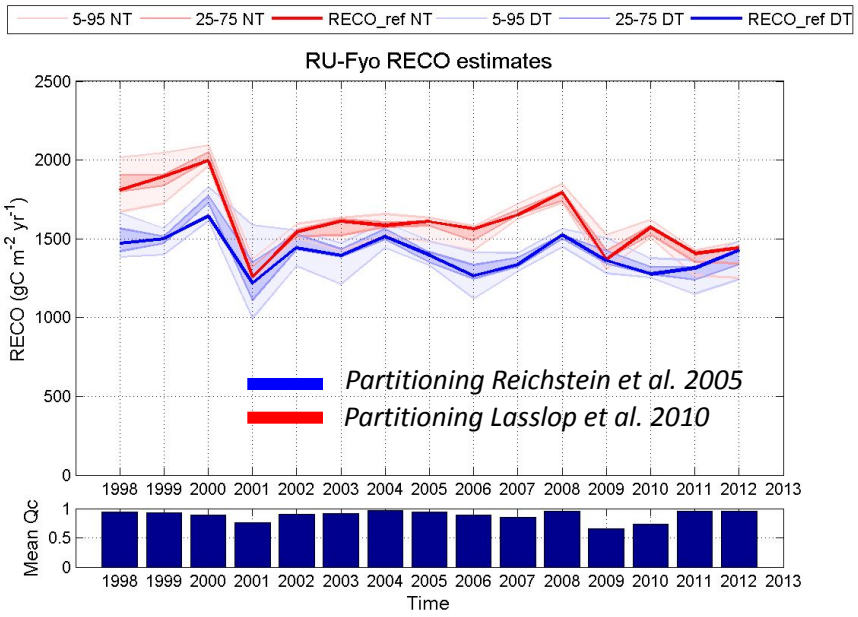
(a non exhaustive list… assuming no errors in measurements)

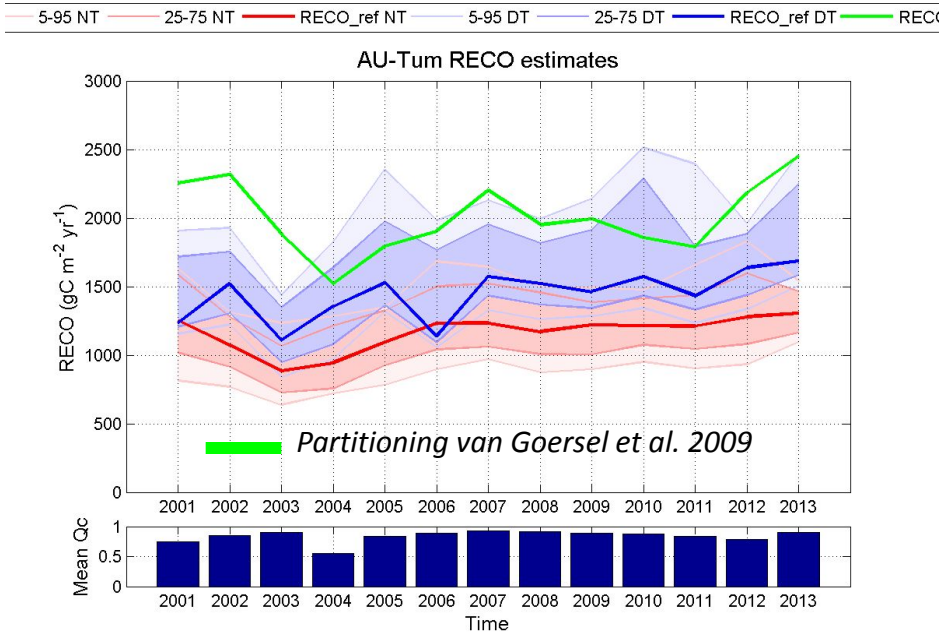
- Where do I put the tower? -> Location (footprint)
- Which height and direction do I put the system? -> footprint
- Which sensors do I use? -> instruments
- How do I collect the data? -> setup
- How do I calculate the fluxes? -> raw data processing
- How do I measure the storage? -> storage
- How much is the random uncertainty? – random uncertainty
- How do I calculate ustar threshold? -> ustar method
- How much is the uncertainty in ustar? -> ustar threshold
- How well ustar filter our advection -> ustar application
- How do I fill the gaps in the data? -> gap filling
- Which method do I use for partitioning? -> partitioning method
- ….

New NEE processing and uncertainty estimation









FINAL REMARKS ON PROCESSING

- Post-processing is important and if not always correctly applied the results could be completely wrong
- Data quality should be always checked carefully also using consistency tests with correlated variables (e.g. Tair, Tsonic, Tsoil, the radiations, Precip and SWC) and looking to the whole time series
- Ustar filtering is a major source of uncertainty and for this reason special attention needs to be used when applied
- Partitioning is a modeling exercise and for this reason also with high uncertainty
- The storage measurement is important. Remember to monitor it at you site (also for other gases…)
- Use of different partitioning methods helps to better understand and quantify the uncertainty, in particular the one using different data (daytime and nightime)