A Training Course on CO_2 Eddy Flux Data Analysis and Modeling

Gap Filling: Theory

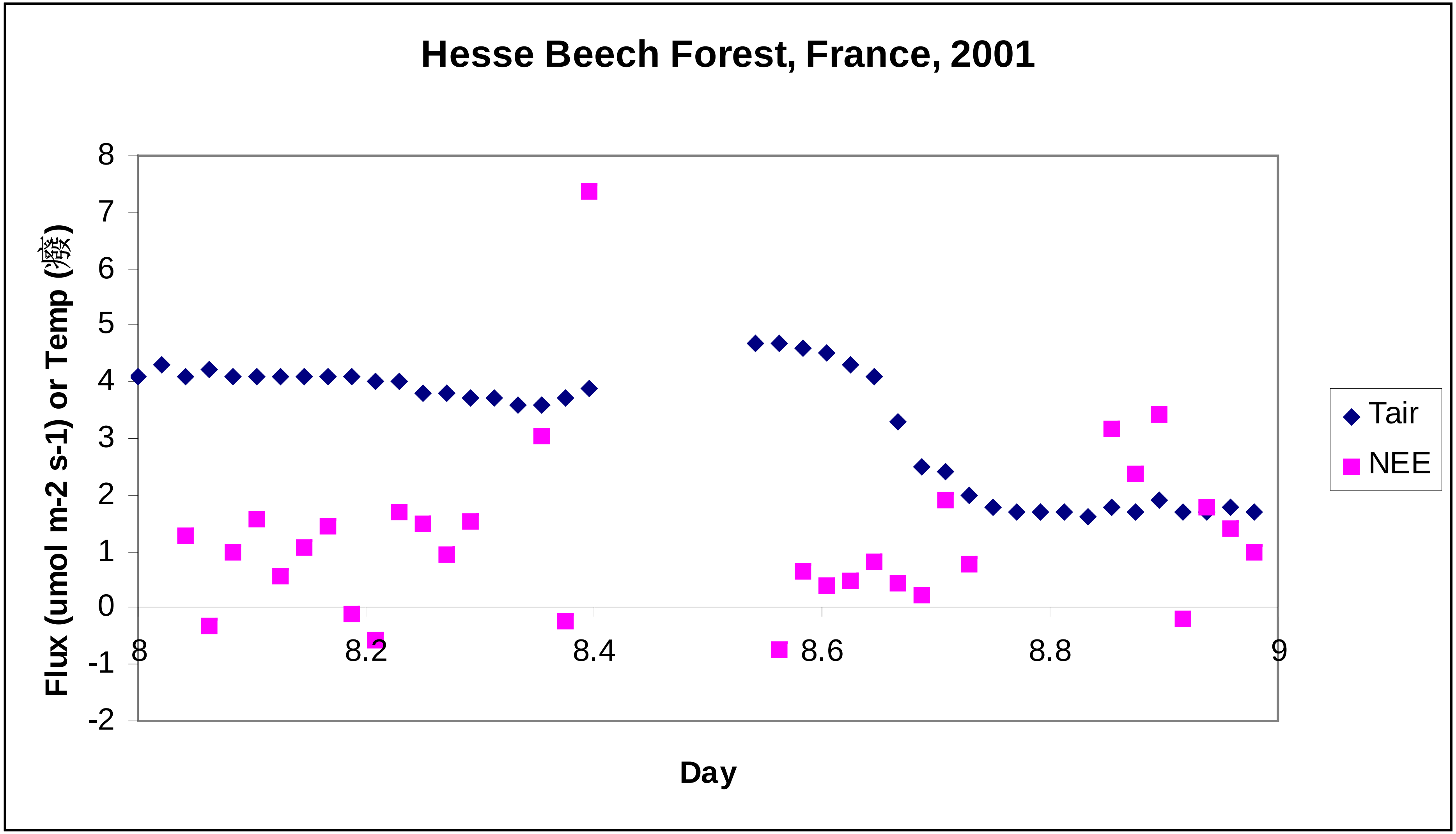
Katherine Owen
John Tenhunen
Xiangming Xiao

Institute of Geography and Natural Resources, Chinese Academy of Sciences, Beijing, China
Institute for the Study of Earth, Oceans and Space, University of New Hampshire, USA
Department of Plant Ecology, University of Bayreuth, Germany

The Institute of Geography and Natural Resources, CAS, Beijing, China

July 24, 2006

Data Gaps


Hesse Beech Forest, France, 2001
Flux (umol m-2 s-1) or Temp (°C)
8
7
6
5
4
3
2
1
0
-1
-2
8
8.2
8.4
8.6
8.8
9
Day
Tair
NEE

Theory: Why gap fill data?

- Stakeholders (i.e., funding agencies, government environmental agencies, policy makers, etc.) want complete annual sums of fluxes
- Modelers want complete time series for coarse time step models

What causes / should cause data gaps?

Site Dependent:

- System or sensor breakdown, calibration, maintenance
- Spikes in raw data, vertical angle of the wind vector
- Wind direction (tower influence, patchiness, meandering footprint area)
- Precipitation and high humidity limits for open path sensors
- Farming farming or management activities

**Remember:* Data gaps due to system failure or rejection of data are not random*

What causes / should cause data gaps?

For All Eddy Covariance Towers:

- Low turbulence (nighttime - u* filtering)
- Quality flags (i.e., Foken and Wichura, 1996) -
  - Steady State tests (covariance of the measured vertical wind and horizontal wind component) ST_{cov}
  - Integral Turbulence Characteristics test- (turbulent conditions using flux-variance similarity) ITC_{σ}

stationarity test ST_{Cov} (deviation in %)	integral turbulence characteristic ITC_{σ} (deviation in %)	QC- flag
< 30	< 30	**1**
< 100	< 100	**2**
> 100	> 100	**3**

Description of the classes:
Class 1: high quality data, use in fundamental research possible
Class 2: moderate quality data, no restrictions for use in long term observation programs
Class 3: low data quality, gap filling necessary

**Remember: Data gaps due to system failure or rejection of data are not random*

What causes / should cause data gaps or recalculation of fluxes?

For All Eddy Covariance Towers:

- Correction for storage
  - calculated from CO_2 profile measurements
  - negligible in short vegetation
  - sums to zero over longer time integrals (d-m-y))
- Correction for advection - evident at sites with 'complex' terrain indicated e.g. by:
  - mean vertical velocity with diurnal pattern (e.g. negative at night, positive during the day)
  - wind direction shifts downhill during nocturnal hours
  - horizontal CO_2 gradients inside vegetation during calm periods

***Remember: Data gaps due to system failure or rejection of data are not random*

Workshop on Gap Filling Comparison
September 18-20, 2006, Jena, Germany Organizer: Antje Moffat

Workshop Goals

- Overview of current gap filling techniques for eddy covariance data
- Statistical evaluation of the different techniques
- Establishment of credibility and reliability for stakeholders and modelers: daily and annual sum with uncertainties
- Proposal for a standardized methodology for CarboEurope IP processing

Gap Filling Technique	Member	Abbrev.
:::::::::: Data-based non-linear regressions ::::::::::		
Non-linear Regression (AQRTa model)	Askoo Noormets	NLR_A
Non-linear Regression (Eyring, Michaelis-Menten (ER,GEP))	Ankur Desai	NLR_EM
Non-linear Regression (2nd order Fourier, Michaelis-Menten) OLS = Ordinary-Least-Squares, AD = Absolute-Deviation	Andrew Richardson	NLR_FM
Non-linear Regression (Lloyd+Taylor, Michaelis-Menten)	Eva Falge	NLR_LM
Non-linear Regression (empirical ER, GEP) FCRN - Fluxnet Canada Research Network	Alan Barr	NLR_FCRN
:::::::::: Artificial neural networks ::::::::::		
Artificial Neural Networks	Dario Papale	ANN1
Artificial Neural Networks	Antje Moffat	ANN2
Baysian Regularized ANN with time series filtering	Rob Braswell	BRANN
:::::::::: Other ::::::::::		
Multiple Imputation Method	Dafeng Hui	MIM
Mean Diurnal Variation	Eva Falge	MDV
Look-Up Tables	Eva Falge	LUT
Marginal Distribution Sampling	Markus Reichstein	MDS
Semi-Parametric Light-Use Model	Vanessa Stauch	SPM
Dual Unscented Kalman Filter (Lloyd+Taylor, Michaelis-Menten)	Dave Hollinger, Jeff Gove	UKF_LM
BETHY, a process-based model	Jens Kattge	BETHY

Gap filling should …

- Rely as much as possible on the data and little as little as possible on external (model theory) assumptions
- Provide unbiased estimates
- Best approximate the statistical properties of the stochastic process (i.e. simulate the individual measured data)

OR

- Best approximate the expected value of the flux

Types of Gap Filling Techniques

time-autocorrelations and/or meteorological constraints methods

- **Empirical** - mean diurnal variation, look up tables (PPFD, Tair, VPD), nonlinear regressions (PPFD, Tair) (Falge et al. AFM 2001). Use priori knowledge to create functional relationships for constraining gap-filled predictions (e.g. NEE ~ Rg)
  - pre-assumed relationships can bias gap-filled predictions as the exact form of functional relationships can be ambiguous
- **Statistical** - Artificial Neural Networks (Papale & Valentini, GCB 2003), Multiple Imputation (Hui et al. AFM 2004), State Dependent Parameter Estimation (Jarvis et al. GCB 2004)
  - using poorly sampled eddy covariance data alone as constraints can bias predictions

Gap-filling: general

	Method	Relies on… / Exploits….
Simple	Linear Interpolation	Linear Interpolation (<1-2 hr gap)
	Site specific ratio between variables	Relationship of variables (i.e. VPD & Tair)
	Near-by meteorological measurements	Near-by weather stations without data gaps
Empirical	Mean diurnal variation (Falge et al. 2001)	Temporal Autocorrelation, diurnal variation
	Non-linear regression (Falge et al. 2001)	Functional dependence on meteo conditions
	Look-up table (Falge et al. 2001)	Dependence on meteo conidtions
Statistical	Neural networks (Papale, Valentini, 2003)	Functional dependence on meteo conditions and time of the year
	Advanced statistical filtering techniques (Multiple imputation (Hui et al. 2004); State dependent parameter estimation, Jarvis et al. 2004)	Functional dependence on meteo conditions, temporal autocorrelation; statistical assumptions (normalty of data etc.)

Simple: Linear Interpolation

- Only use for short gaps (1-3 missing measurements)
- Best for missing meteorological values (Tair, rH, etc.)

Drawbacks:

- Only acceptable for occasional short gaps

(Falge et al. 2001 Agricultural & Forest Meteorology)

Simple: Site specific ratio between variables

- Site specific ratio/equation between
  - PPFD and Rg
  - VPD, Tair, rH
  - u*, momentum, Tair, Pair
- If PPFD is missing but all measurements for Rg are present, use ratio (PPFD/Rg) to gap fill PPFD

Drawbacks:

- Only useful for the above relationships

(Falge et al. 2001 Agricultural & Forest Meteorology)

Simple: Near-by meteorological measurements

- Long period of time (months) without a meteorological sensor(s) due to failure
- Near-by meteorology stations, with similar climate, elevation, etc. could be used.
- Please make modelers and future users of the data aware that the gap was filled with near-by data !

Drawbacks:

- Climate, vegetation, etc. at near-by meteo station can be different than at the eddy covariance tower

Empirical: Mean Diurnal Variation

- The missing half-hour observation is replaced by the mean of that time period on adjacent days, usually a window of 7-14 days.
- Can capture non-linearity due to diurnal & temporal changes in response

Drawbacks:

- If gaps are biased towards a condition (cloudy periods, etc.) then the gap filled value will not be representative of the condition
- No functional responses between fluxes and meteo variables (bias on cloudy & sunny days)
- Best with short gaps (<14 days)

(Falge et al. 2001 Agricultural & Forest Meteorology)

Empirical: Nonlinear Regression Methods

- Nighttime NEE data filled using (4 or 6 period/yr):

$$NEE_{night} = \mathrm{Re}\,co_{night} = R_{ref} \cdot e^{E_0\left(\frac{1}{T_{ref}-T_0} - \frac{1}{T_K - T_0}\right)}$$ Lloyd & Taylor

or

$$NEE_{night} = \mathrm{Re}\,co_{night} = R_{ref} \cdot e^{\frac{E_A}{R}\left(\frac{1}{T_{ref}} - \frac{1}{T_K}\right)}$$ Arrhenius

or

$$NEE_{night} = \mathrm{Re}\,co_{night} = A \cdot e^{(B \cdot T)}$$ Van't Hoff

- Daytime NEE data filled using:

$$NEE_{day} = \frac{a' \cdot Q_{PPFD} \cdot F_{GPP,sat}}{F_{GPP,sat} + a' \cdot Q_{PPFD}} - F_{RE,day}$$ Michaelis-Menten

or

$$NEE_{day} = F_{GPP,opt}\left(1 - e^{\frac{a' \cdot Q_{PPFD}}{F_{GPP,opt}}}\right) - F_{RE,day}$$ Misterlich

Drawbacks:

- No inclusion of VPD or water stress
- Management (mowing, harvesting, etc.)

(Falge et al. 2001 Agricultural & Forest Meteorology)

Empirical: Look-up Tables

- Missing values of NEE are looked up in a table based that give mean and s.d. of NEE based on Tair and PPFD conditions
- 4 or 6 seasonal periods per year
- PPFD classes 0, 1-200, 201-400,…, 2001-2200 $\mu mol\ m^{-2}\ s^{-1}$ (Different light response curves possible)
- Tair classes -19 - -17, -17 - -15,…, 47 - 49° C

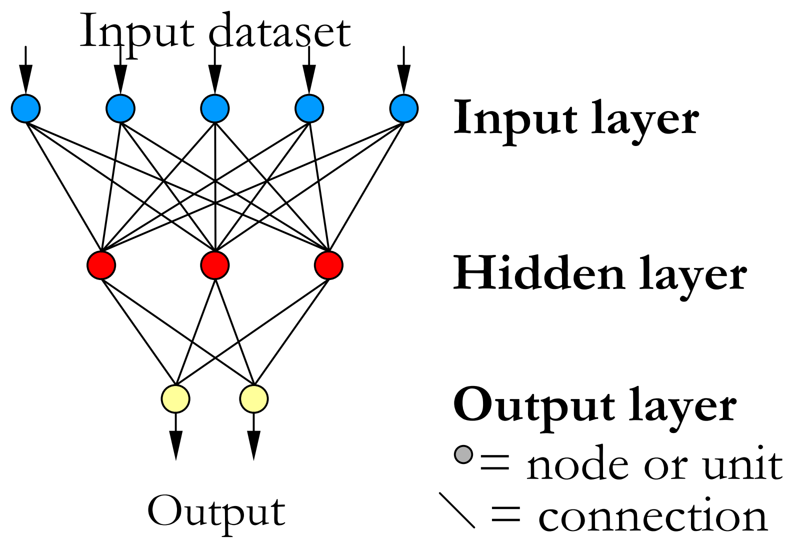
Drawbacks:

- Tair and PPFD data must be available for missing periods
- Scatter in data due to water stress, heterogeniety of fetch/footprint area, etc.

(Falge et al. 2001 Agricultural & Forest Meteorology)

Statistical: Artificial Neural Networks

- A good dataset of real observations are used to train the network -input & outputs known (the connections' weight are set)
- The network is validated on other datasets by choosing the input variables, number of layers & nodes
- Then run for datasets with NEE gaps



Network of many simple processing elements "nodes"

(Papale & Valentini 2003 Global Change Biology)

Statistical: Artificial Neural Networks

- Positive Aspects:
- Observations parameterize and validate the model
- No need to know relationship between inputs and outputs
- Good estimates of long gaps

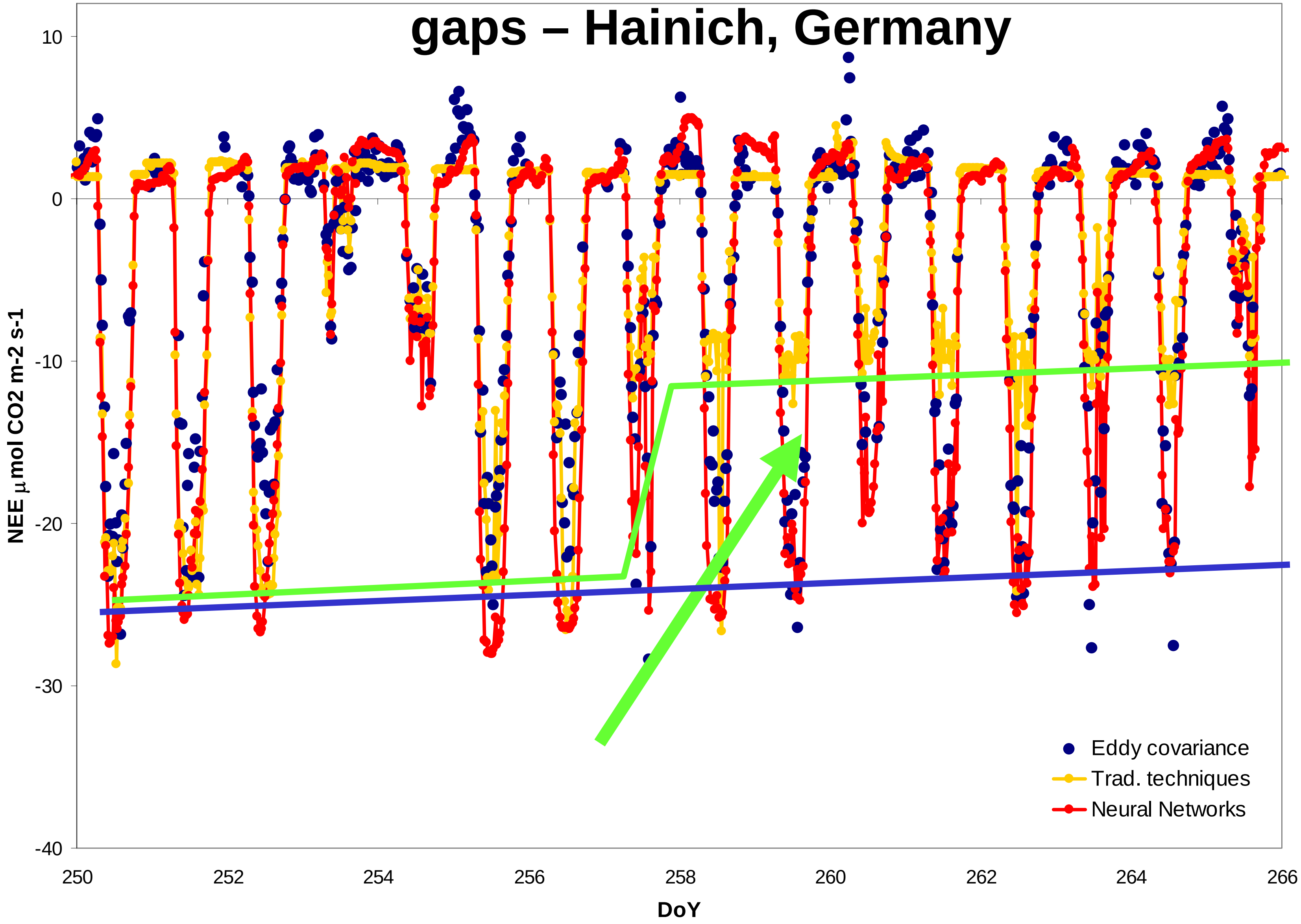
Drawbacks:

- Meteorological gaps must be filled before with another method
- Black box
- Time

(Papale & Valentini 2003 Global Change Biology)

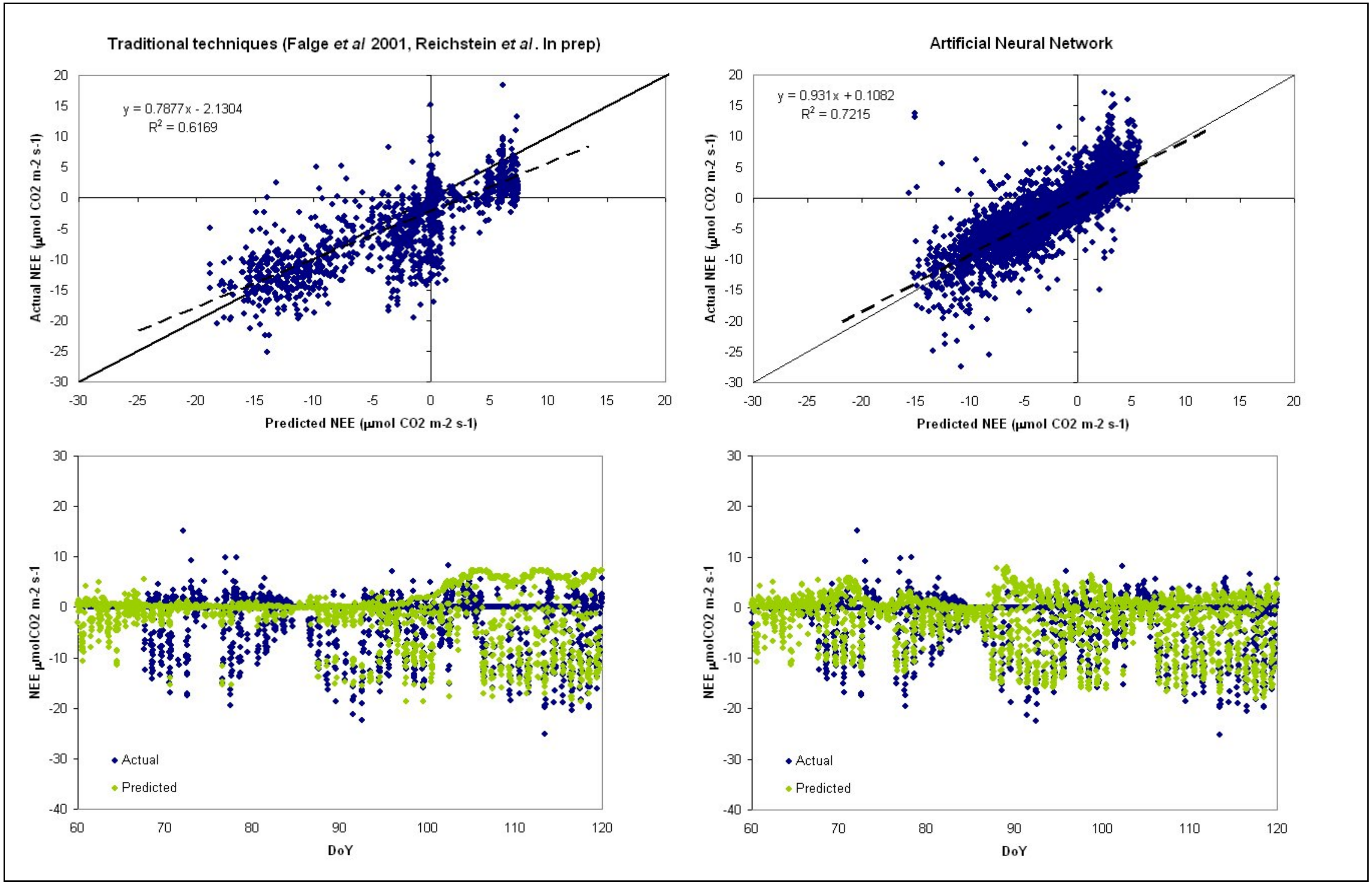
Statistical: Artificial Neural Networks

Comparison of different gap-filling techniques with artificial gaps – Hainich, Germany



Statistical: Artificial Neural Networks

Comparison of techniques with artificial gaps – Tharandt



Statistical: Multiple Imputation

- Monte Carlo technique comparing observed with estimated missing data
- 3-5 imputations are calculated for each missing data point
- Numerous gap filled data sets are created for each site - mean, variance, standard error are calculated

Drawbacks:

- Gap filled range is typically smaller than observed
- Does not fill winter periods well
- Does not preserve short-term relationships between NEE & Meteo

(Hui et al. 2004 Agricultural & Forest Meteorology)

Statistical: State Dependent Parameter Estimation

- Gap filled by vicinity of the missing data to sorted surface temperature groups
- But calculates NEE value by only solar radiation on the random walk process
- Gaussian-like window function within a Kalman filter-regression framework

Drawbacks:

- Only gap fills NEE
- Gaps in Temp or Radiation must be filled with nearby meteo station data
- Does not predict NEE during times of water stress well

(Young 2000 Nonlinear and Nonstationary Signal Processing; Jarvis et al. 2004 Global Change Biology)

Empirical: Marginal Distribution Sampling *method currently used in Bayreuth

Assumptions:

- NEE = NEE(Rg, Tair, VPD, time) + ε
- NEE (Rg, Tair, VPD, time) $\cong$ NEE(Rg+ΔRg, Tair+ΔTair, VPD+ΔVPD, time+Δtime)
- The smaller Δtime and the more environmental constraints available the better

(Reichstein et al. 2005 Global Change Biology)

Empirical: Marginal Distribution Sampling *method currently used in Bayreuth

- General type of approach same as Falge et al. (2001)
- Combination of Mean Diurnal Variation and Look-Up Table methods
- Differences:
  - Dynamic averaging window size (as small as possible ➔ better exploitation of temporal autocorrelation)
  - „Moving“ look-up table (➔ value to be filled always in the center of the class)

(Reichstein et al. 2005 Global Change Biology)

Empirical: Marginal Distribution Sampling

Quality-controlled half-hourly data (storage, ustar,...) **Reichstein et al. Gap Filling Method**

- NEE present ? → Yes → Don‘t fill:
  - No ↓
- Rg, T, VPD, NEE available with |dt| <=7 days → Yes → Fill with average of available values: Filling quality: A
  - Using similar meteo values (within 50 Wm^{-2}, 2.5°C, 5.0 hPa)
  - No ↓
- Rg, T, VPD, NEE available with |dt| <≤ 14 days → Yes → Filling quality: A
  - ↓
- Rg, NEE available with |dt| <≤ 7 days → Yes → Filling quality: A
  - ↓
- NEE available within |dt| ≤ 1 hour → Yes → Filling quality: A
  - ↓
- NEE available within |dt| <≤ 1 day (& same hour of day) → Yes → Filling quality: B
  - ↓
- Rg, T, VPD available with |dt| <≤ 21, 28, ..., 140 days → Yes → Filling quality: B, if |dt| <=28, else C
  - ↓
- Rg, NEE available with |dt| <≤ 14, 21, ..., 140 days → Yes → Filling quality: B, if |dt| <=14, else C
  - ↓
- NEE available within |dt| <≤ 7, 14, ... days → Yes → Filling quality: C

(Reichstein et al. 2005 Global Change Biology)

Theory: Conclusions

- No world-wide standard method exists, but consensus is forming.

- Different methods may be suitable for different sites

- Different methods may be suitable for different goals & modelling exercises

- In order to model a wide variety sites consistently, we chose Marginal Distribution Sampling (Reichstein et al. 2005)

- Difference in gap filling can cause different annual sums of CO_2 balances

Acknowledgements

- Presentations from Gap filling Workshop - June 9-10, 2004 in Viterbo, Italy

Gap Filling: Practice

- As the gap filling and flux partitioning programs are combined, all practice will be done at the end of the flux partitioning section