Meterological Time series Imputation using Kalman Filters

Simone Massaro - Bioclim seminar 18 Jan 2023

Outline

1. Introduction
2. Kalman Filter
3. Preliminary results
4. Next steps

1. Introduction

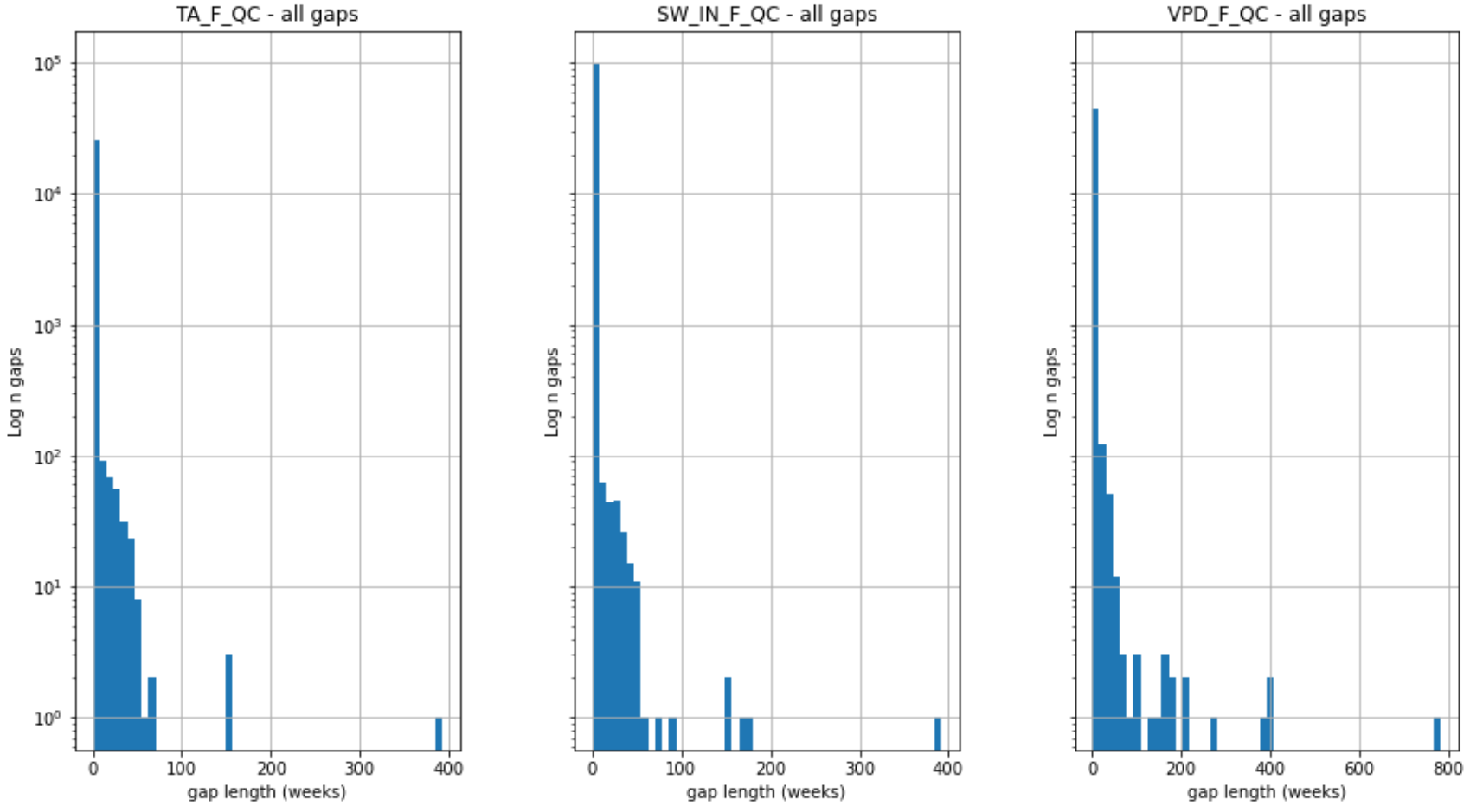
Background

- EC tower measures also meteorological variables (eg. air temperature, wind speed)
- technical issues (eg. broken sensor) result in **meteo time series** with **gaps**
- Presence of gaps is a problem in many EC data applications (eg. ecosystem modelling)

Dataset

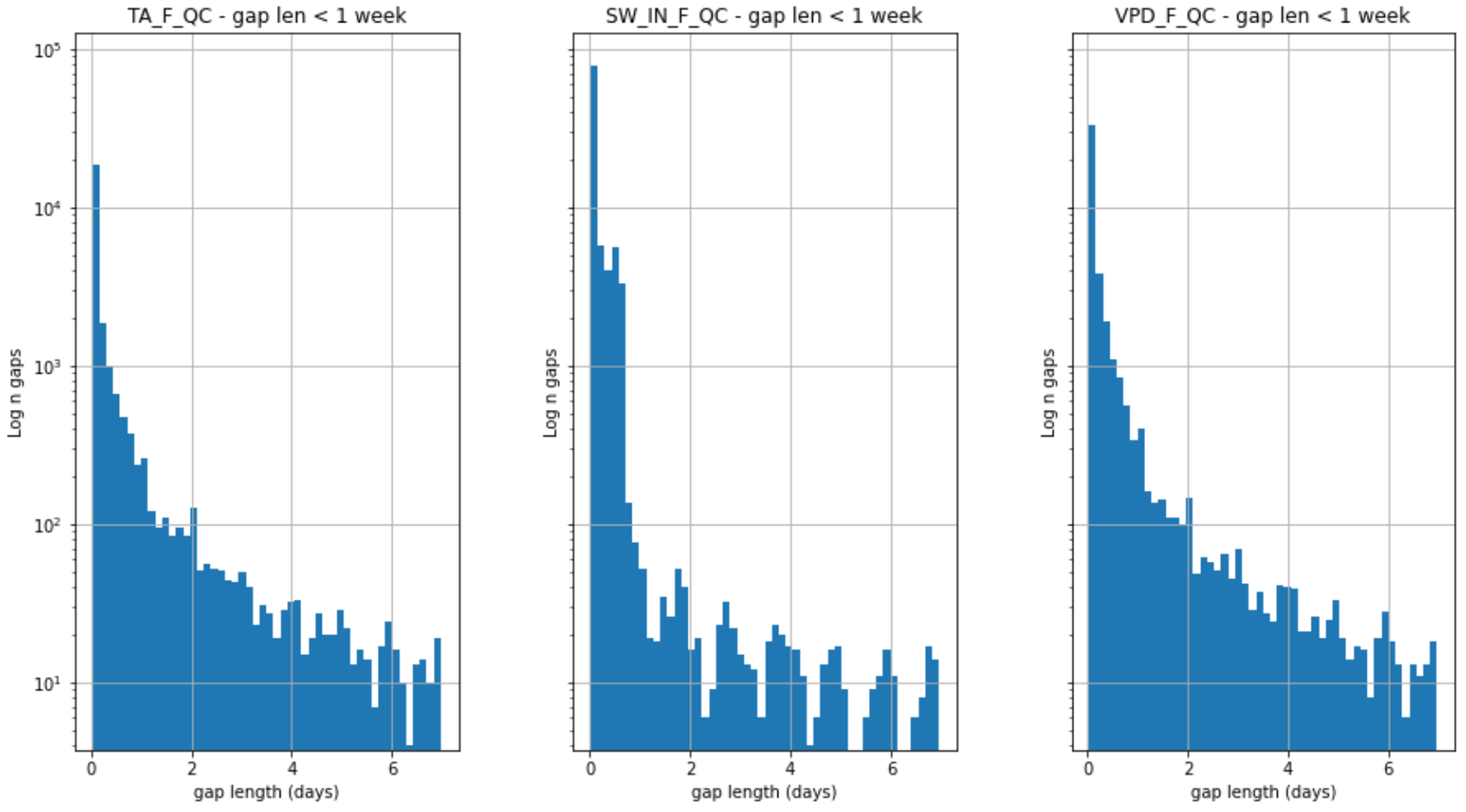
- Fluxnet 2015 data from Hainich (more 20 years)
- Global meteo dataset (downscaled ERA-Interim from Fluxnet 2015)
- meteorological measurements every 30 mins
- focusing on 3 variables
  - Air temperature: `TA`
  - Incoming shortwave radiation: `SW_IN`
  - Vapour Pressure Deficit: `VPD`

Gap len distribution in Fluxnet



plot of distribution of gaps for all `TA`, `SW_IN` and `VPD` for all sites

Gap len distribution in Fluxnet



plot of distribution of small gaps (<200) for all TA, SW_IN and VPD for all sites

How to fill gaps

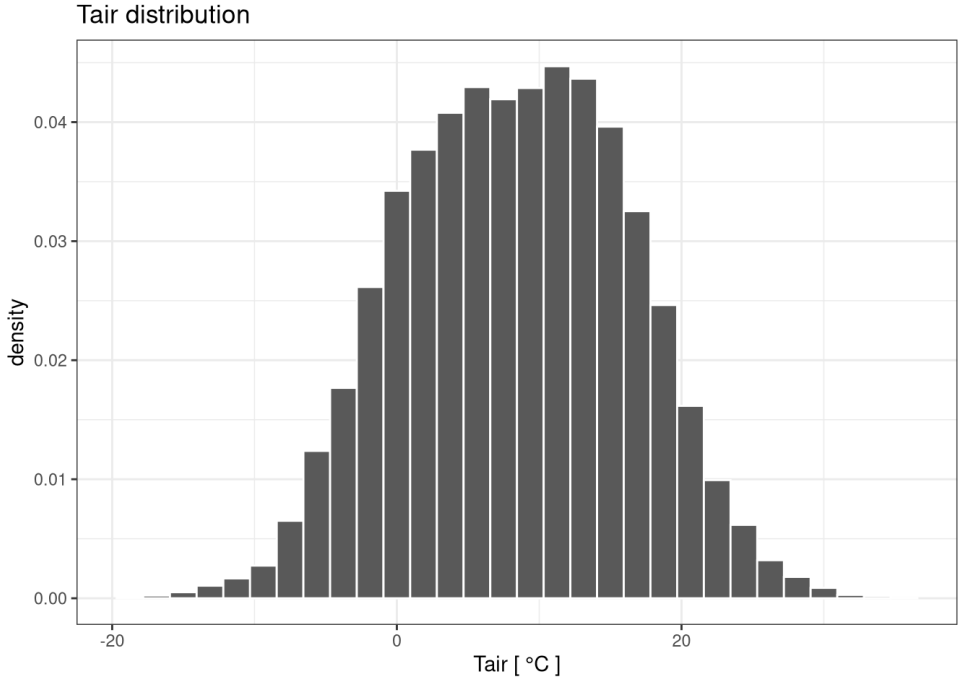
1. use previous and following measurements for one variable and temporal auto-correlation (eg. diurnal cycles)
2. correlation with other variables measures (eg. solar radiation and temperature)
3. other measurements of meteo variables (eg. nearby station)

State of the art

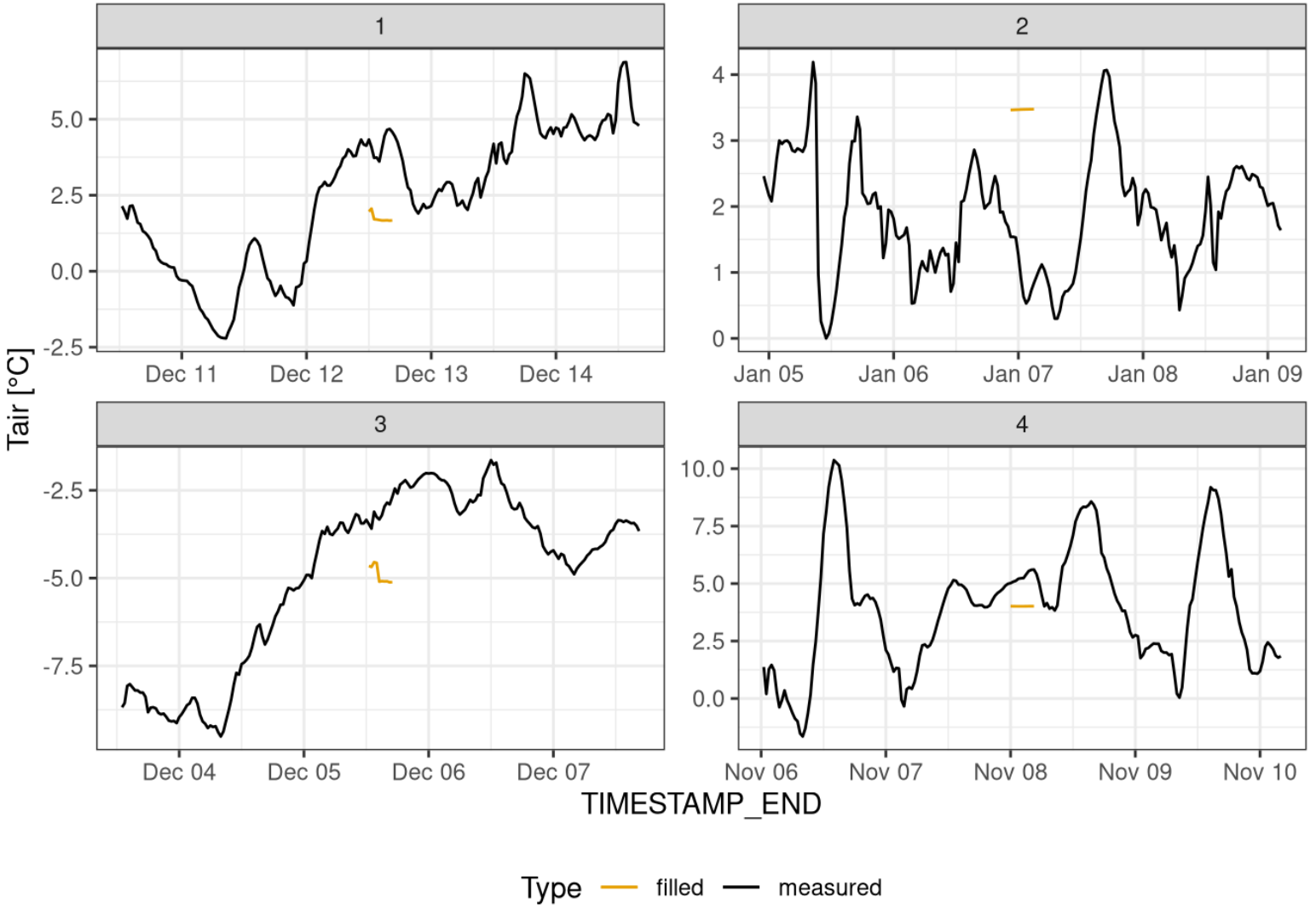
- OneFlux pipeline (Fluxnet + ICOS + AmeriFlux)
- Short and medium gaps using Marginal Distribution Sampling (MDS)
- Long gaps filled with ERA data (global meteo dataset) using linear regression to reduce site bias

How MDS (Marginal Distribution Sampling) works

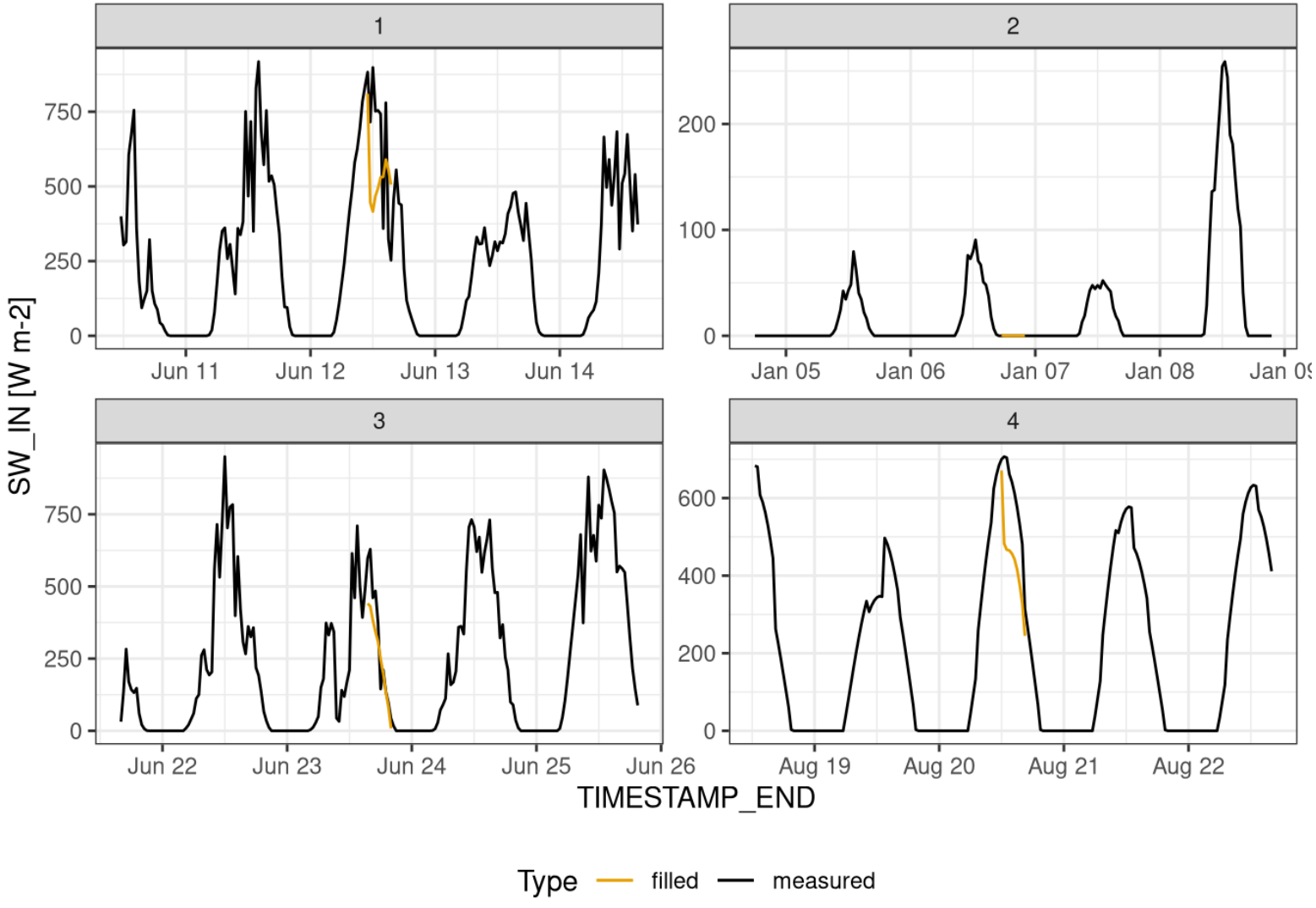
- take a time window (7 days) around the gap
- use 3 predictors variables (TA, SW_IN and VPD) and divide them in n discrete bins
- for each bin (combination of conditions) find the average value of the missing variable
- for each gap find the closest condition and fill with the average value
- if necessary increase the time window
- quality flag depends on the time window size



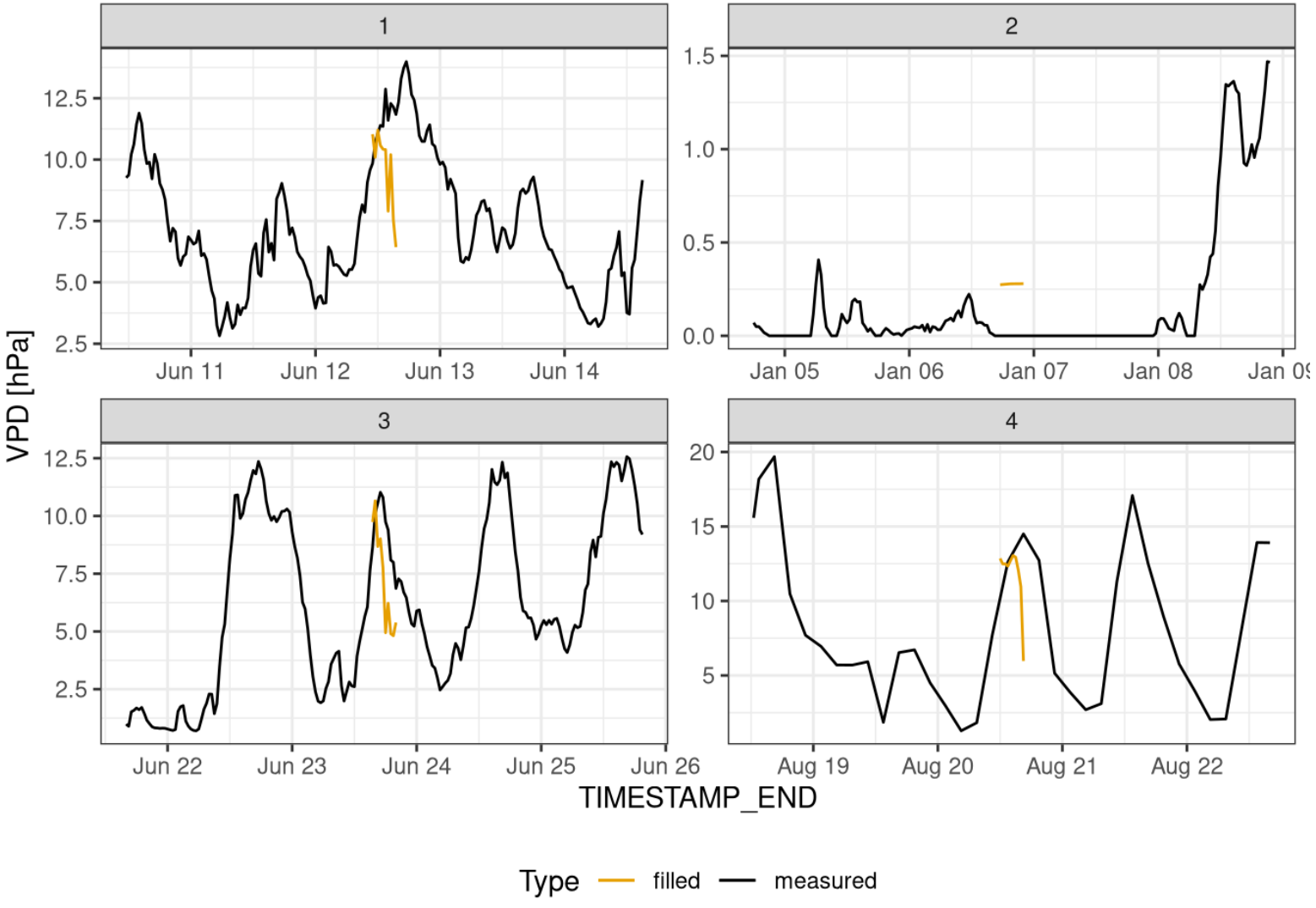
MDS - gap filling TA



MDS - gap filling SW_IN



MDS - gap filling VPD



MDS limitations

- No uncertainty for the predictions (only a quality flag)
- Don’t combine the different imputation approaches
- high error for medium/short gaps

Thesis goal

- develop model to impute missing data in meteorological time series
- include all 3 imputation approaches
- provide an uncertainty of the predictions

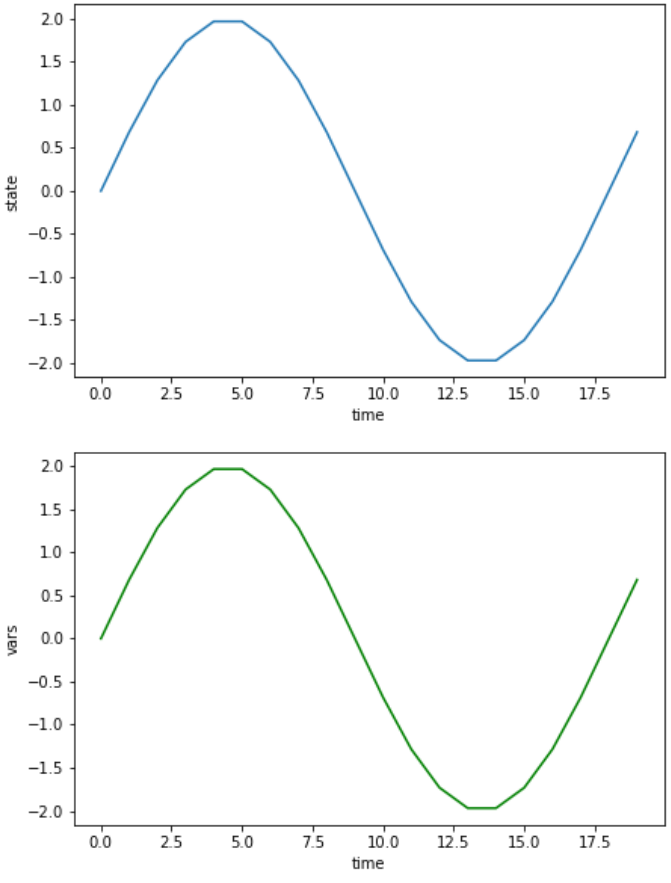
3. Model: Kalman Filter

How Kalman Filter works

Models over time a **latent** variable (we are not observing it), the **state** of the system.

The current state x_t depends using:

1. the previous state x_{t-1}
2. current observation y_t
3. control variable c_t (ERA data)

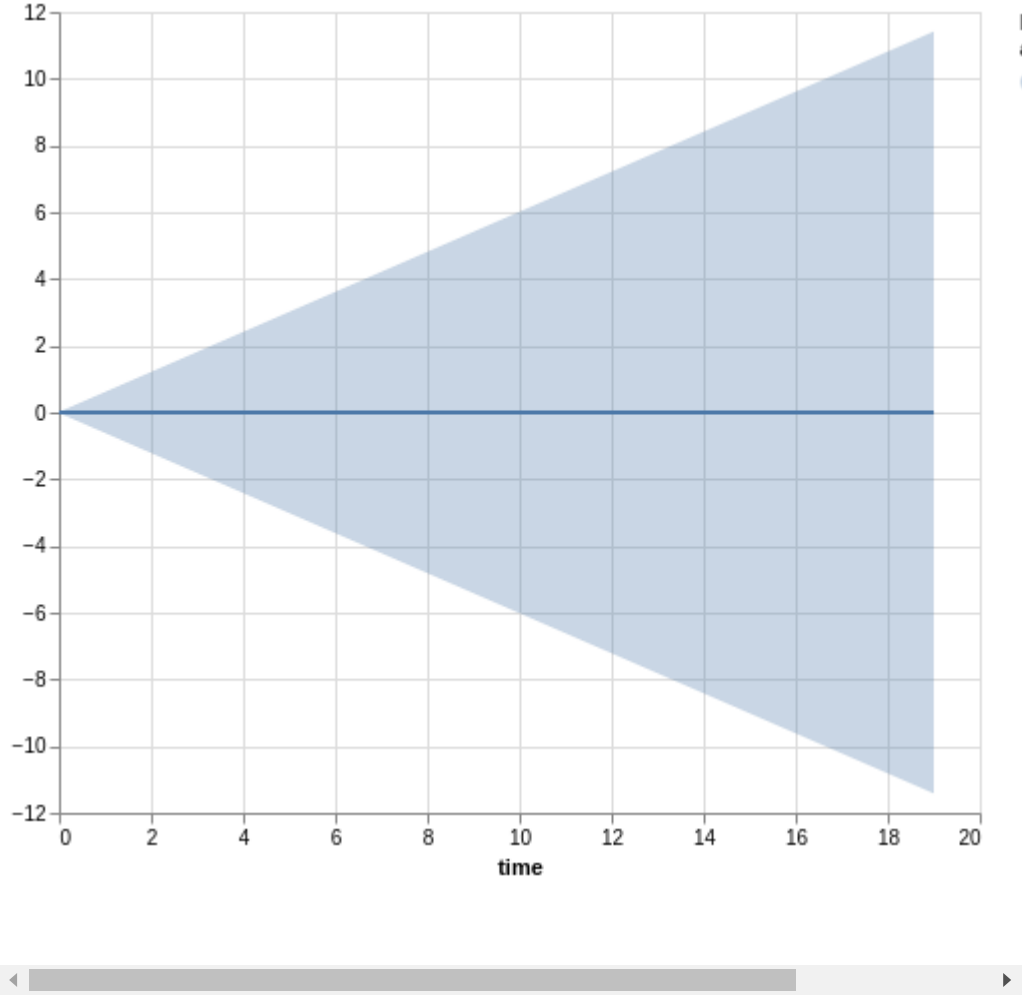


1. Previous state

$$x_t = Ax_{t-1} + \varepsilon$$

where:

- x_t is the current state
- x_{t-1} is the previous state
- A is a linear transformation of x_{t-1}
- ε is the “process” noise which is a random variable with a normal distribution with mean 0



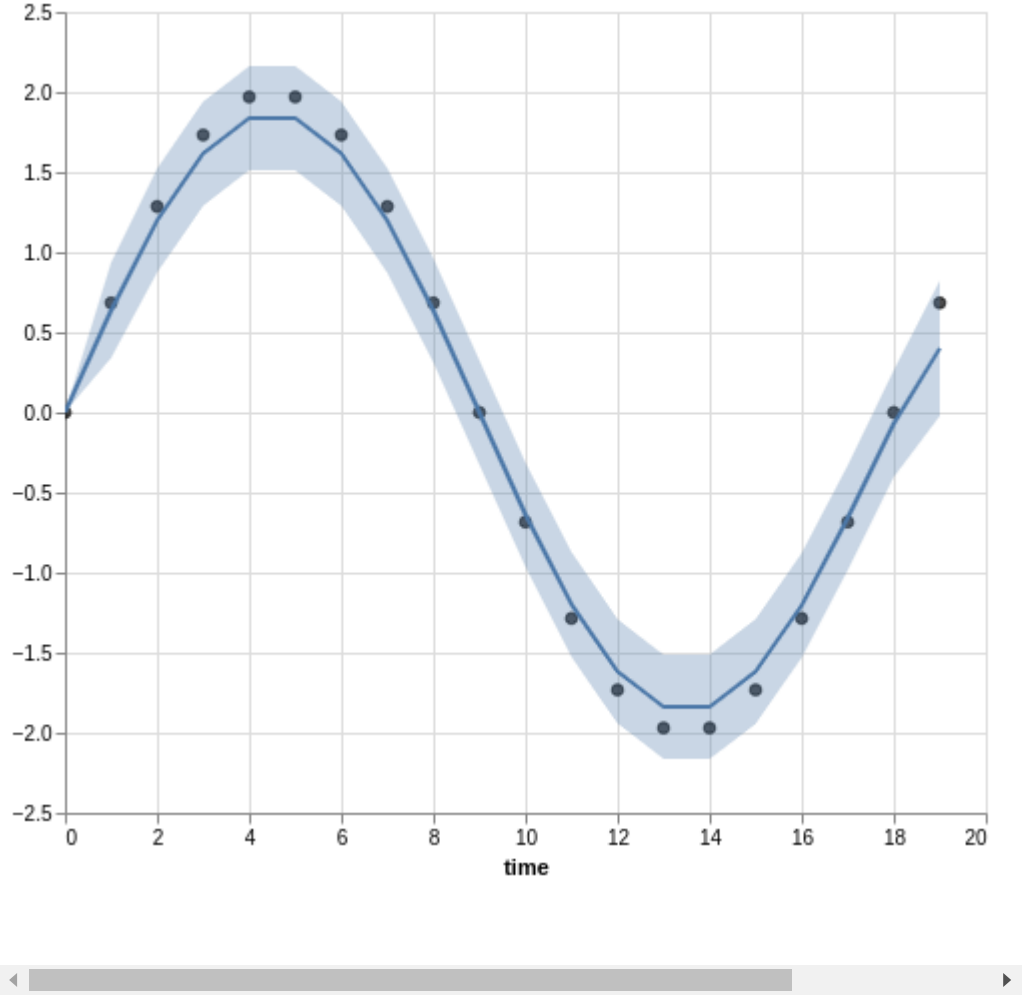
2. Current Observation

$$x_t = Ax_{t-1} + \varepsilon$$
$$y_t = Hx_t + \nu$$

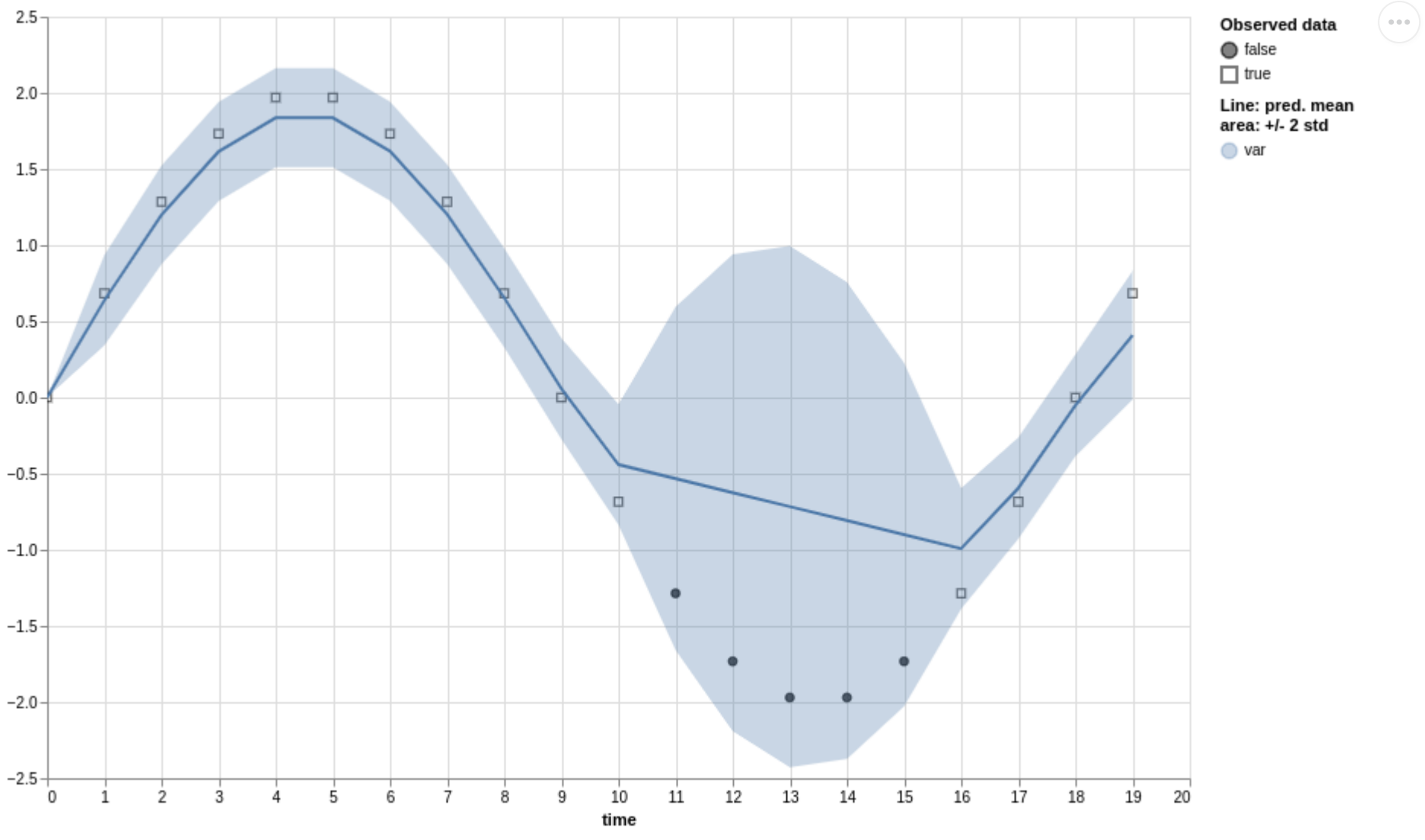
where:

- y_t is the current observation
- H is a linear transformation of y_t
- ν is the "observation" noise which is a random variable with a normal distribution with mean 0

using the rules of probabilistic inference if we observe y_t you can update the distribution of x_t



Gaps


2.5
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2.5
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
time
Observed data
false
true
Line: pred. mean
area: +/- 2 std
var

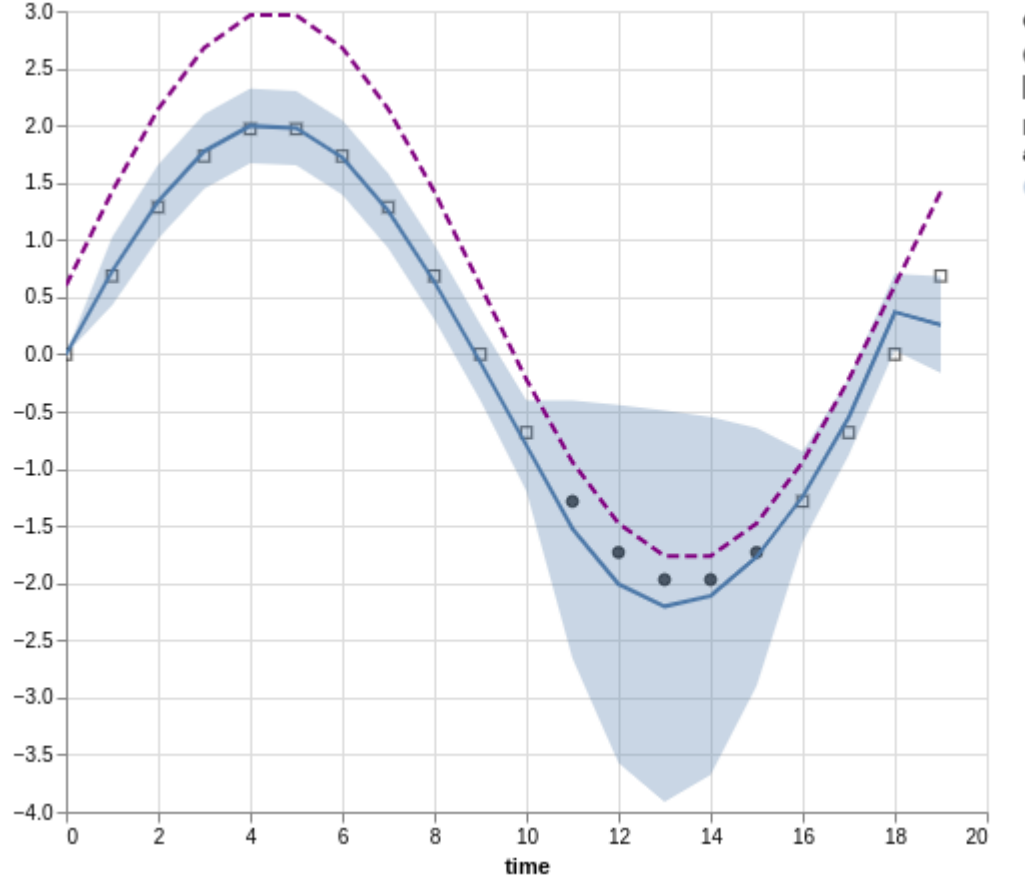
3. Control variable

$$x_t = Ax_{t-1} + Bc_t + \varepsilon$$
$$y_t = Hx_t + \nu$$

where:

- B is a linear transformation of c_t

Use the difference between current and previous value of control variable



Extra: Variable correlation

How to find model parameters

- create artificial gaps
- predicting gap in the model
- compute the log likelihood of the predictions
- maximise the log likelihood

Kalman Filter

pros:

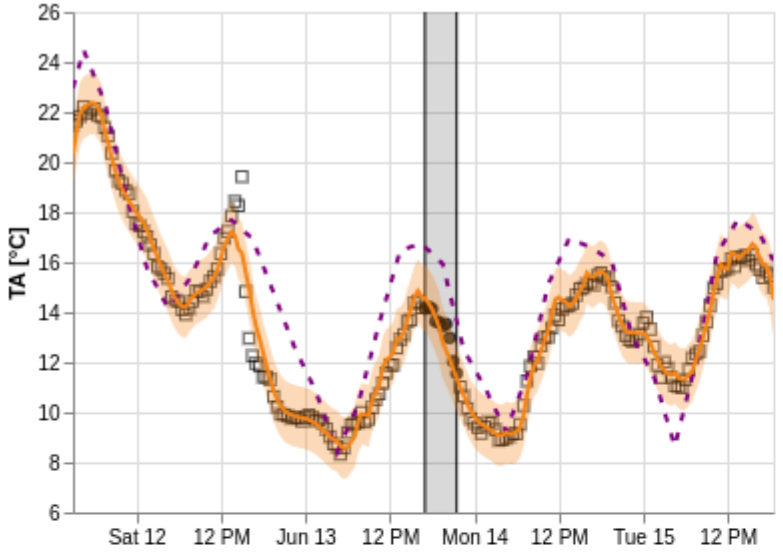
- Probabilist model: the output of the model is a **distribution of predictions**, not a single value
- Combines **all** 3 approaches to gap filling in one model
- interpretable paramters
- computationally efficient

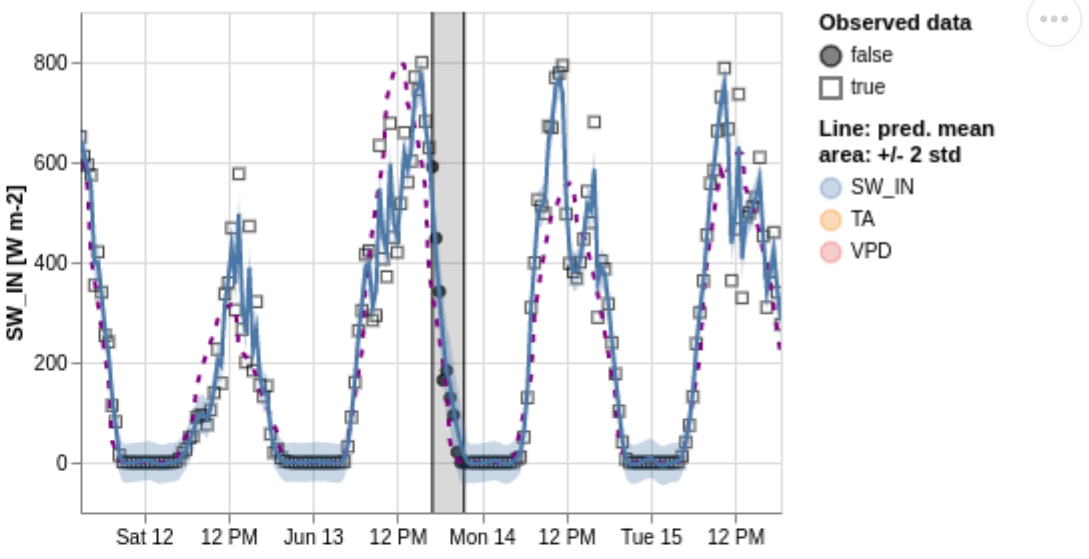
cons:

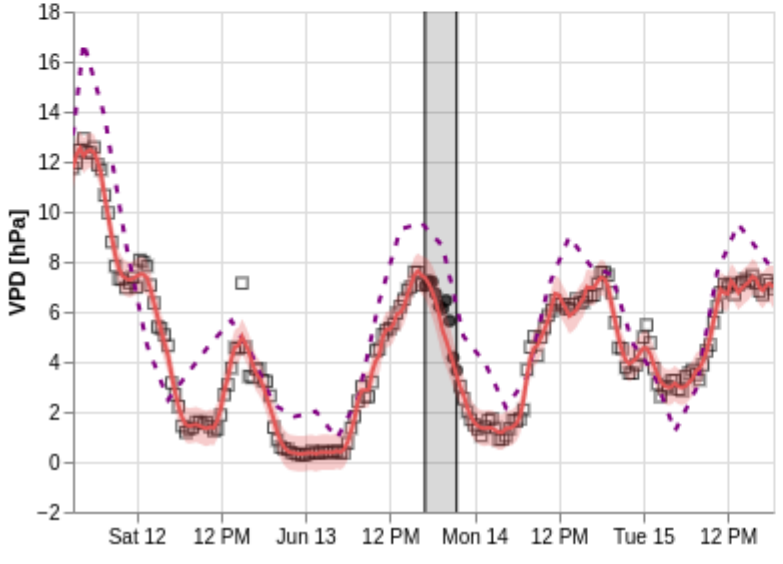
- keeps tracks only of the local state

2. Preliminary results

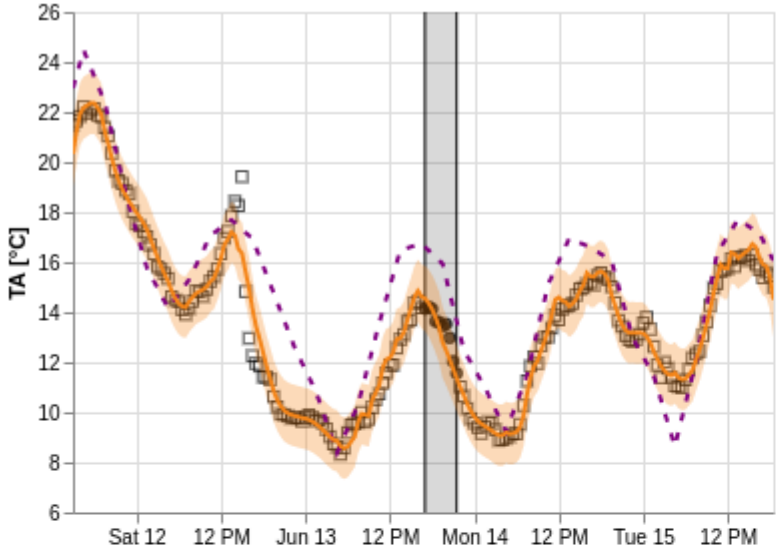
Kalman Filters gap #1

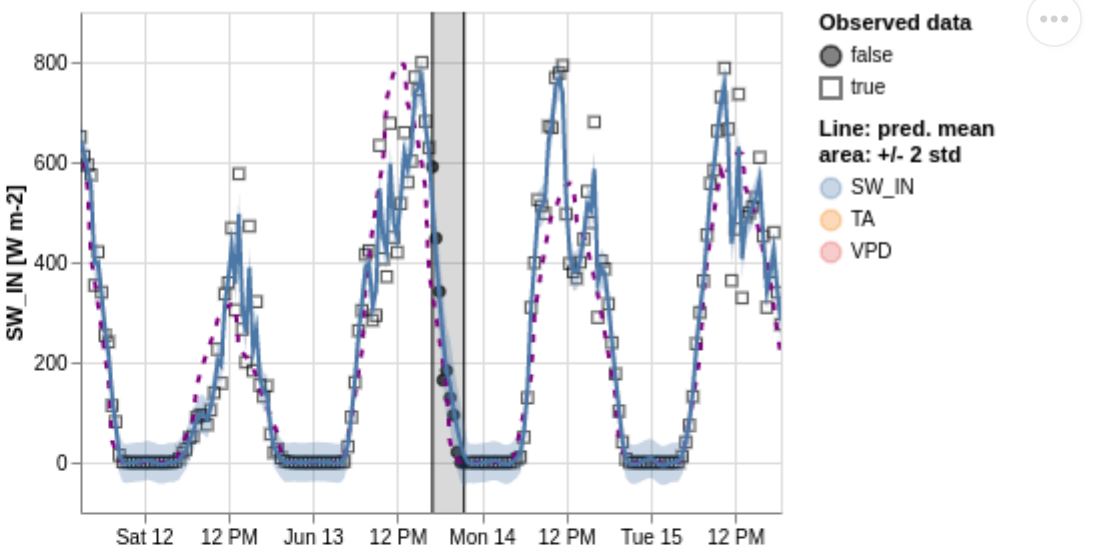


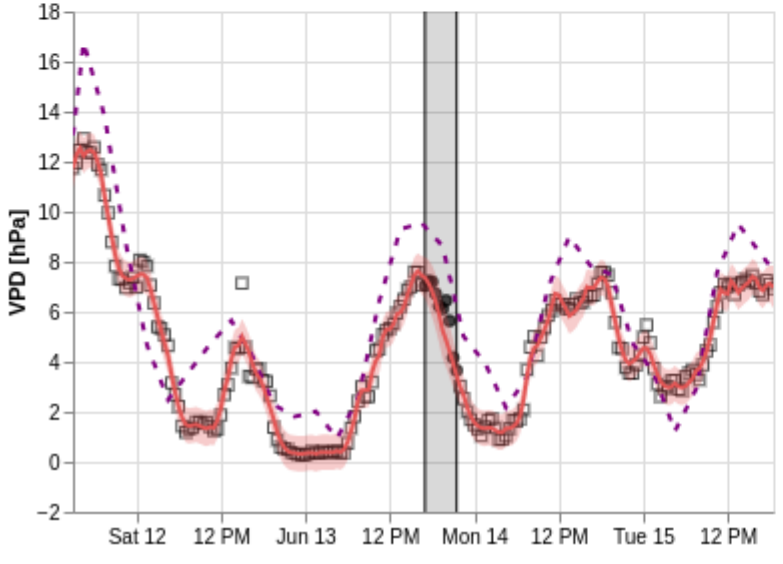




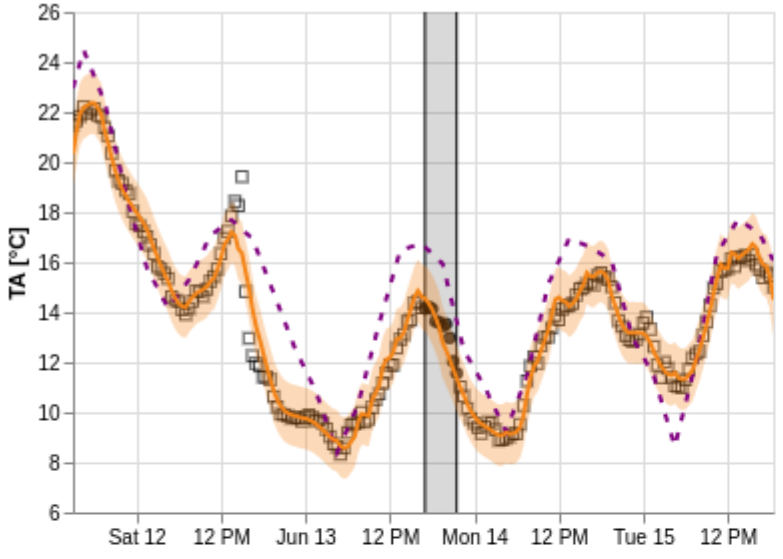
Kalman Filters gap #2

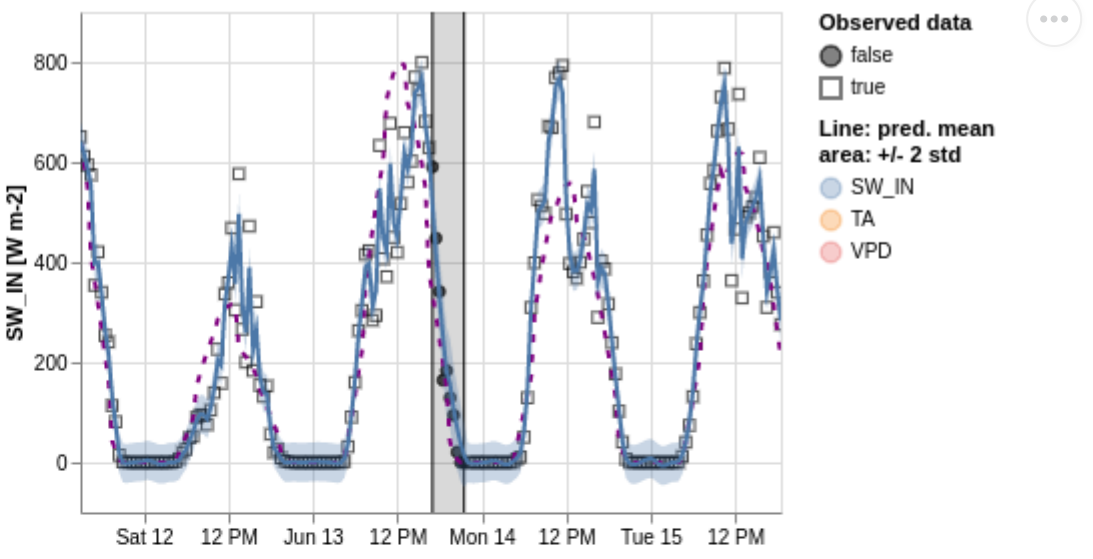


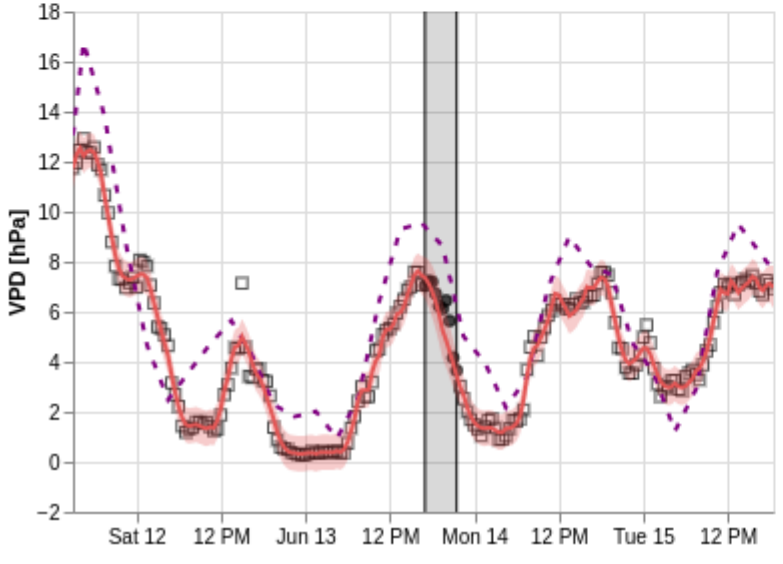




Kalman Filters gap #3







Next steps

What is missing in the model development

- improve numerical stability of model (work in progress)
- find optimal settings for training and inference
  - n observations before after/gap
  - how to best generate artificial gaps

How to assess the model?

Open questions:

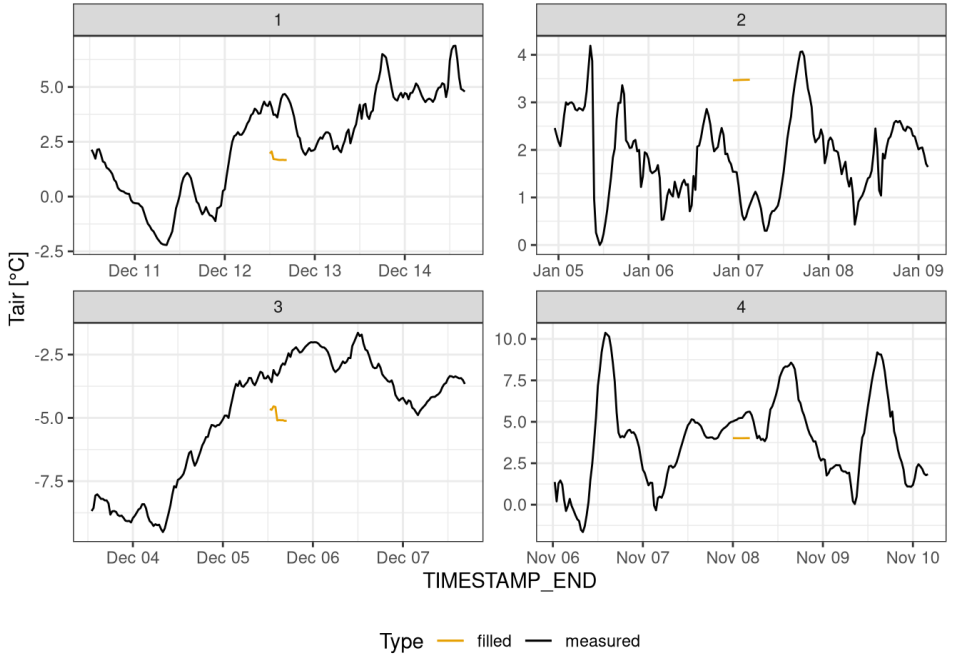
- how to choose gap lengths?
- how to choose number of variables missing?
- which variable to focus on?

How to assess the model? Metrics

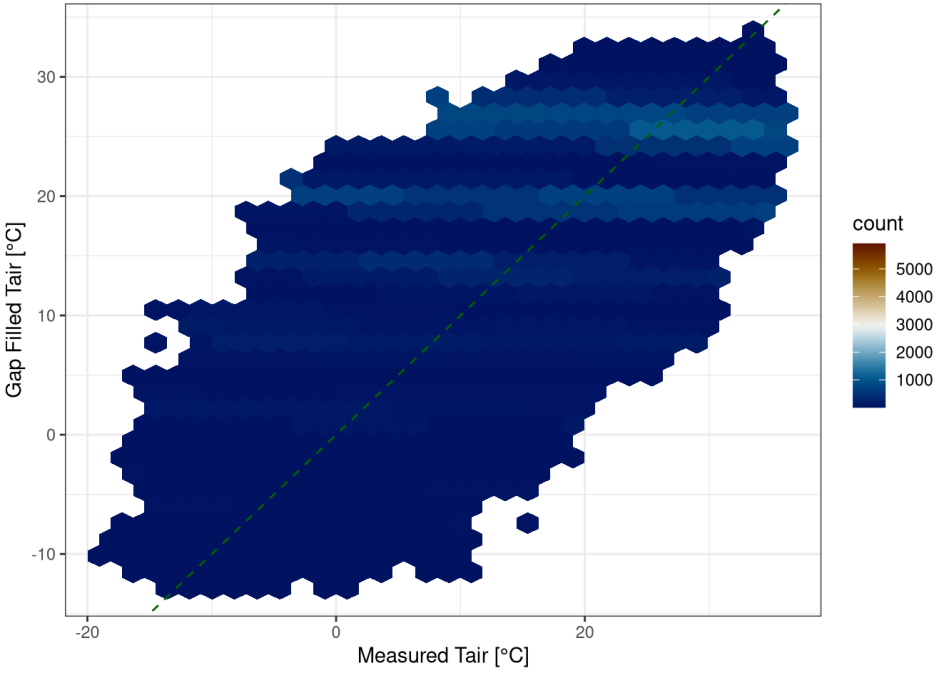
- RMSE - interpretation difficult as it’s relative to the variable
- r2 - gaps are often too short to interpret properly
- ?

How to assess the model? Figures

Time series

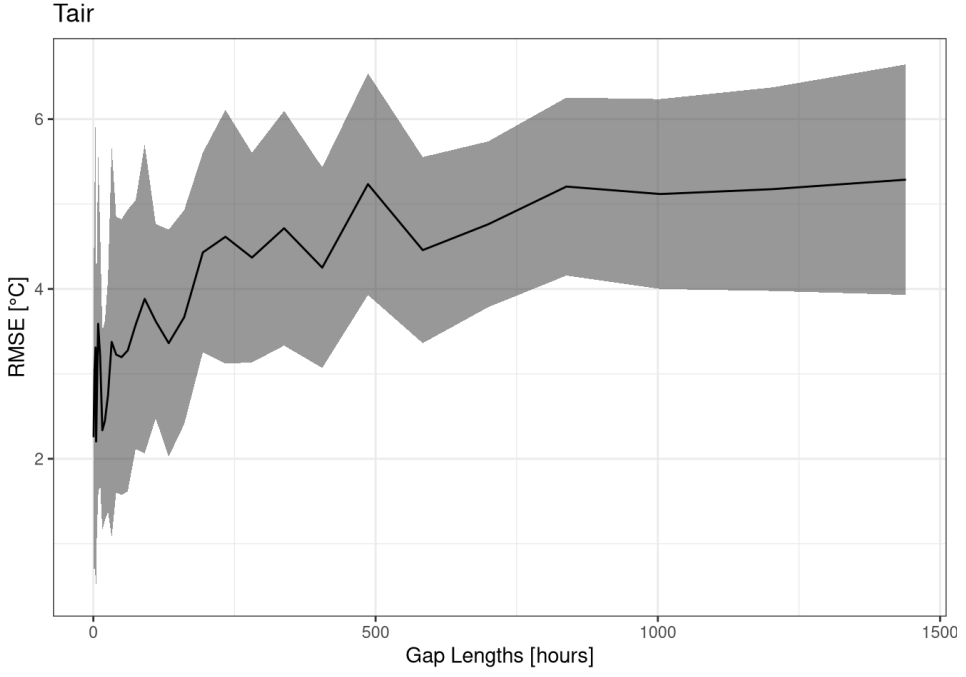


Scatter plots

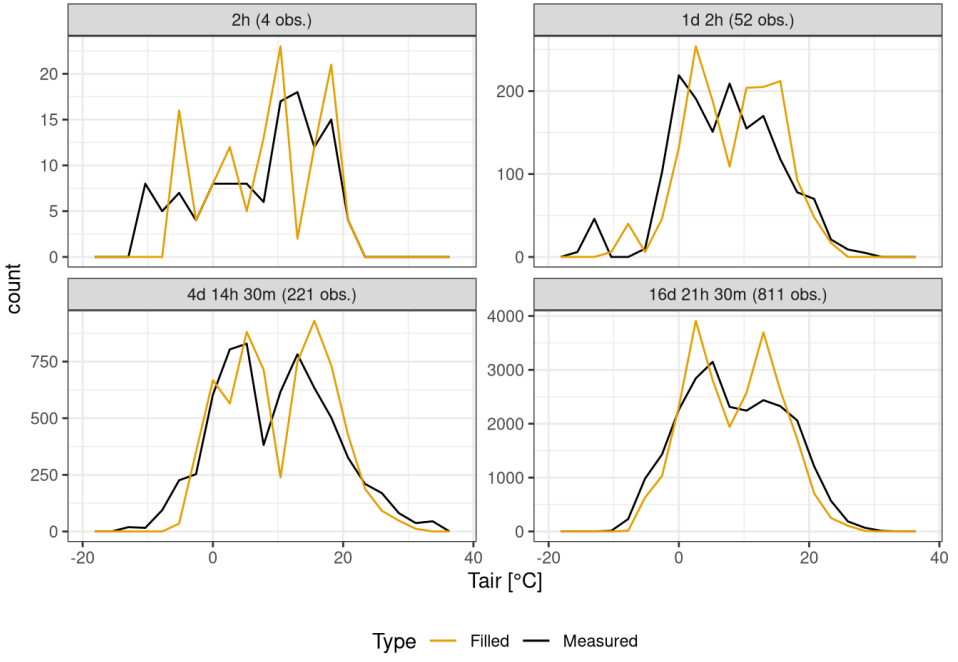


How to assess the model? Figures

Gap length / mean RMSE



Distribution gaps vs filled



Model use

- what is the importance of the model performance?
  - can optimize kalman filter performance
- what is the impact of better gap filling for data users?
  - why better filling for short gaps is useful
  - how can the uncertainty be used

Future outlook

- provide pre-trained model on Fluxnet 2015 and then to fine-tune to local site
- provide web-service for filling gaps
- reprocess Fluxnet 2015 dataset

Questions & Comments