



**Barcelona
Supercomputing
Center**

Centro Nacional de Supercomputación



**EXCELENCIA
SEVERO
OCHOA**

R tools user meeting

An-Chi Ho and Núria Pérez-Zanón

contributors: Jaume Ramon

05/03/2021

Agenda

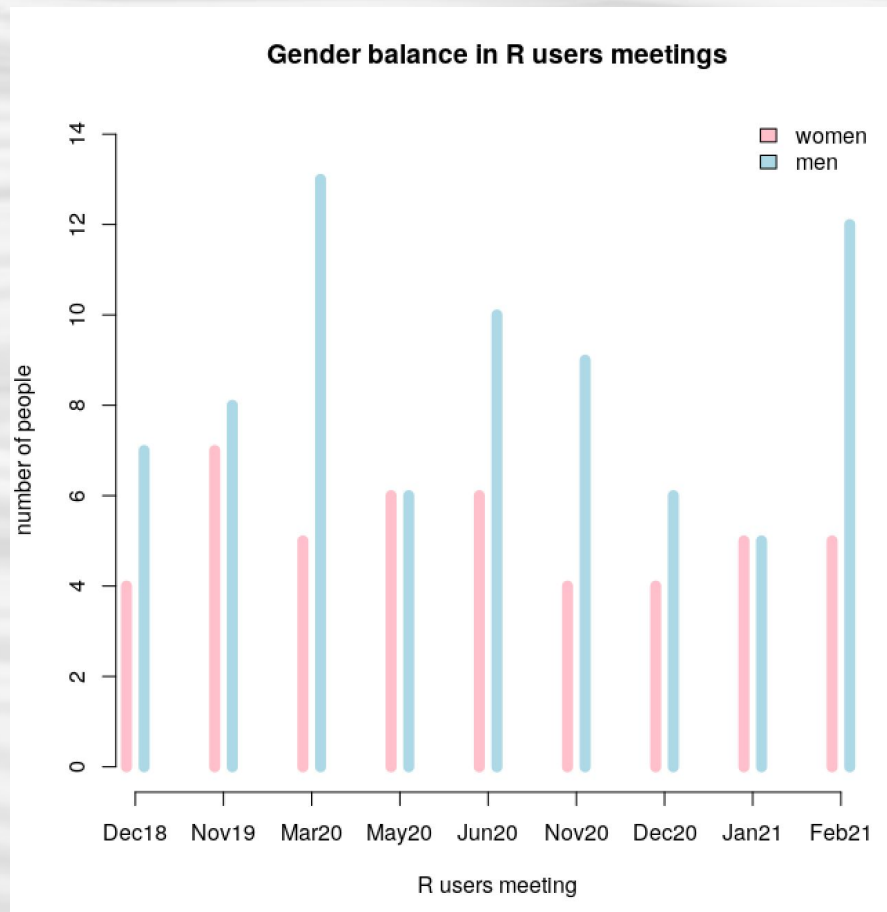
1. Package update
 - CTools
 - startR
2. Quality Control of wind series in R: a guided tour (Jaume)
3. Q&A

Before start ...

Given the WomenES initiative, we have used the information from our minutes to create this plot (which is not planned to be presented anywhere else for now).

As expected, given the balance gender in the department, here, generally, we are less women than men.

I have never thought that we can do anything wrong during the meetings about discrimination (in all the meanings of discrimination) but please, if you feel uncomfortable in any sense, please, say it during the meeting, tell us later or our talk to somebody else if needed.



CSTools



**New CSTools
4.0.0**

- available on CRAN:

<https://CRAN.R-project.org/package=CSTools>

- extended vignettes section

Analogs

Achieving Best Estimate Index

Data Storage and Retrieval

Ensemble Clustering

Most Likely Terciles

Multi-model Skill Assessment

Multivariate RMSE

Plot Forecast PDFs

RainFARM

Weather Regime Analysis

- installed on workstations and Nord 3

Basic functions

CST_Load
CST_Anomaly
CST_SaveExp
CST_SplitDim
CST_MergeDims
s2dv_cube
as.s2dv_cube

Correction

CST_BiasCorrection
CST_Calibration
CST_QuantileMapping
CST_BEI_Weighting
BEI_PDFBest
CST_DynamicalBC

Downscaling

CST_Analogs
CST_RFTemp
CST_RainFARM
CST_RFSlope
CST_RFWeights
CST_ADAMONT
CST_AnalogsPredictors

Evaluation

CST_MultivarRMSE
CST_MultiMetric

Plotting functions

PlotMostLikelyQuantileMap
PlotForecastPDF PlotPDFsOLE
PlotCombinedMap PlotTriangles4Categories

Classification

CST_WeatherRegimes CST_RegimeAssign
CST_CategoricalForecast
CST_EnsClustering CST_MultiEOF

 To be included

 Enhancements for December

 New method

startR

startR 2.1.0-3

- Installed in WS and Nord3, only R \geq 3.6.1.
- Bugfixes:
 - If the name of `return_vars` is the synonym of inner dim name, change it back to inner dim name ([issue 87](#))
 - Remove incorrect check of character selector (e.g., region) ([issue 91](#))

```
obs <- Start(  
  dat = repos_obs,  
  var = 'tos',  
  time = 'all',  
  lat = values(list(-10, 10)),  
  lat_reorder = Sort(),  
  lon = values(list(10, 20)),  
  lon_reorder = CircularSort(0, 360),  
  synonyms = list(lat = c('lat', 'latitude'),  
                  lon = c('lon',  
  'longitude')),  
  return_vars = list(latitude = NULL,  
                    longitude = NULL,  
                    time = 'date'))
```

BAD

```
obs <- Start(  
  dat = repos_obs,  
  var = 'tos',  
  time = 'all',  
  lat = values(list(-10, 10)),  
  lat_reorder = Sort(),  
  lon = values(list(10, 20)),  
  lon_reorder = CircularSort(0, 360),  
  synonyms = list(lat = c('lat', 'latitude'),  
                  lon = c('lon', 'longitude')),  
  return_vars = list(lat = NULL,  
                    lon = NULL,  
                    time = 'date'))
```

GOOD

startR -- wrapper of Start()

The wrapper of Start(), multiStart(), is coming soon...

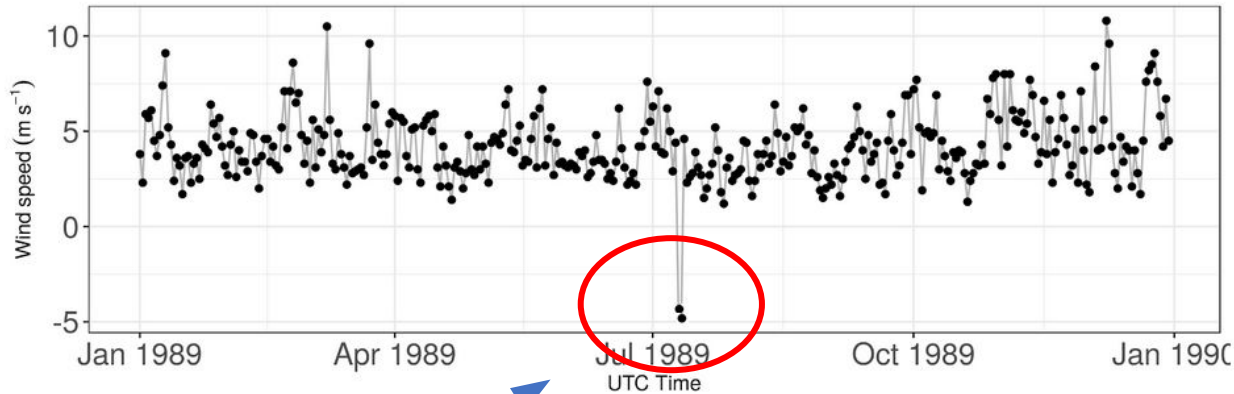
- For multiple datasets that have different calendars/sdates/ftime
- Assign different selector values for each dataset

```
time_hadgem3 <- c(60, NA, 61) # 1230, NA, 0101
time_mpi_esm <- c(60, 61, 62) # 1230, 1231, 0101

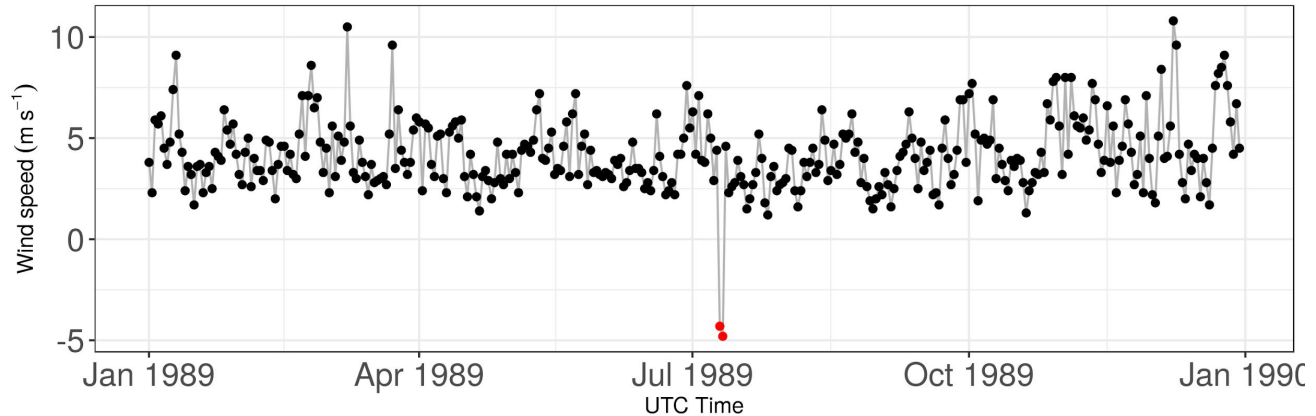
res <- multiStart(
  dat = list(list(name = 'hadgem3', path = path_hadgem3),
             list(name = 'mpi_esm', path = path_mpi_esm)),
  var = 'tasmax', sdate = '2000',
  fyear = list(list(name = 'hadgem3', fyear = fyear_hadgem3),
              list(name = 'mpi_esm', fyear = fyear_mpi_esm)),
  time = list(list(name = 'hadgem3', time = time_hadgem3),
             list(name = 'mpi_esm', time = time_mpi_esm)),
  lat = indices(3:5), lon = indices(1),
  return_vars = list(lat = NULL, lon = NULL, time = 'dat'),
  retrieve = FALSE)
```

Quality Control of wind series in R: a guided tour

'strangeR things' in wind speed series



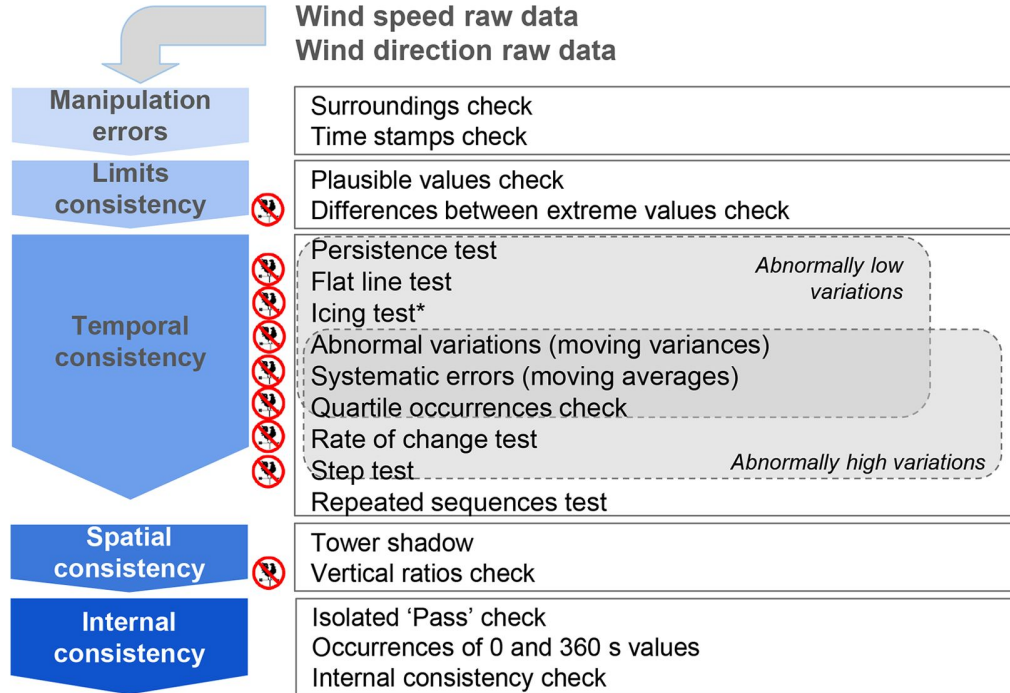
Negative winds!





18 Quality Control tests

to ensure the high quality of wind series



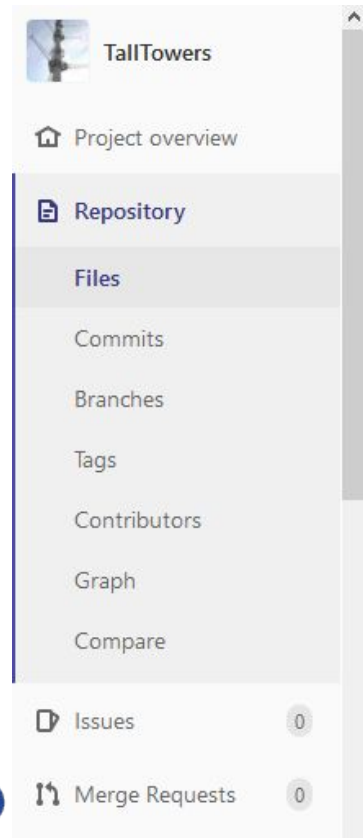
Wind speed QC data
Wind direction QC data



Not applicable to wind direction data

(*) Needs temperature data

The code in R runs the QCs sequentially



TallTowers

- Project overview
- Repository**
- Files
- Commits
- Branches
- Tags
- Contributors
- Graph
- Compare

Issues 0

Merge Requests 0

```
109
110 #-----
111 # QC TESTS
112 #-----
113 print(paste0("Starts at ",Sys.time()))
114
115 all.data <- pausable_values(all.data) # Default: max.value.fail=113.2,max.value.suspect=75
116 print(paste0("Finished plausible values at ",Sys.time()))
117
118 all.data <- flat_line(all.data)
119 print(paste0("Finished flat line at ",Sys.time()))
120
121 all.data <- step_test(all.data)
122 print(paste0("Finished step test at ",Sys.time()))
123
124 all.data <- persistence_test(all.data)
125 print(paste0("Finished persistence test at ",Sys.time()))
126
127 all.data <- diff_extreme_values(all.data)
128 print(paste0("Finished difference extreme values at ",Sys.time()))
129
130 all.data <- rate_change(all.data)
131 print(paste0("Finished rate change at ",Sys.time()))
132
133 all.data <- repeated_sequences(all.data)
134 print(paste0("Finished repeated sequences at ",Sys.time()))
135
```

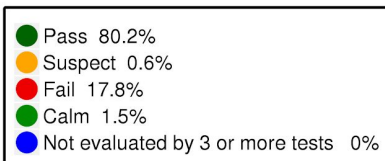
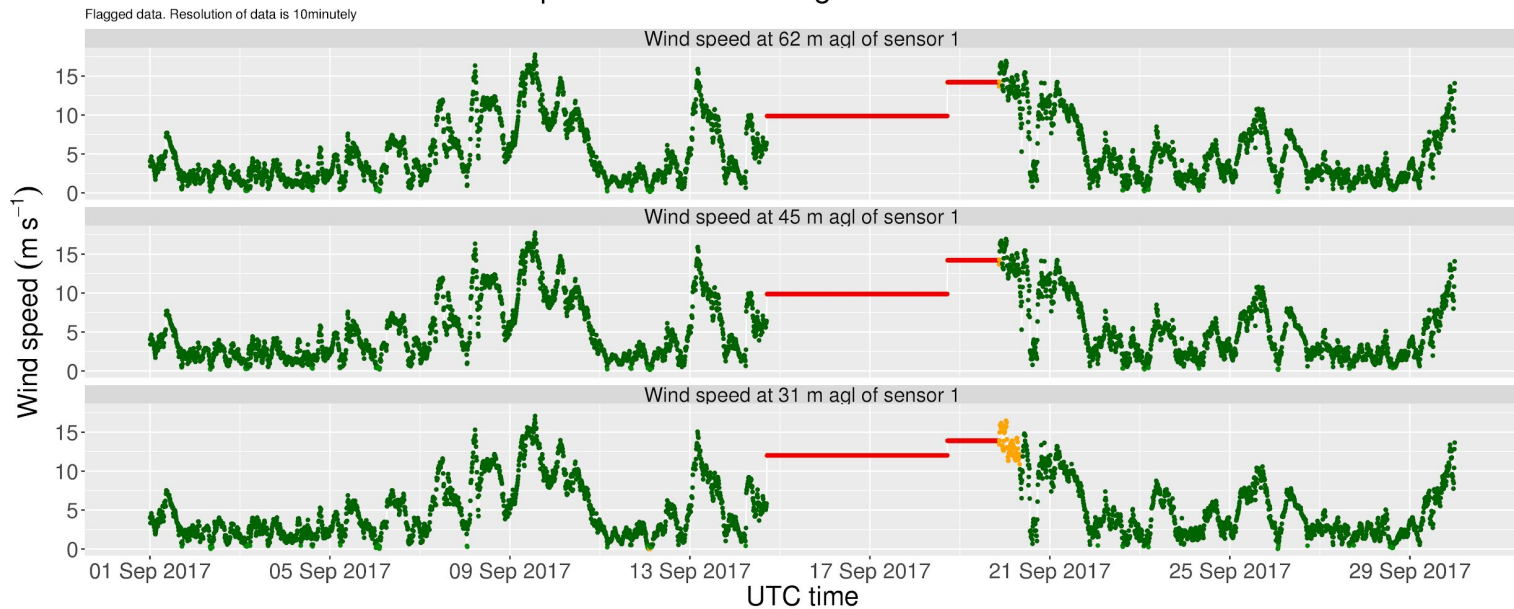


WHAT TYPE OF BAD DATA CAN THE QCs DETECT?

Identical values in a row

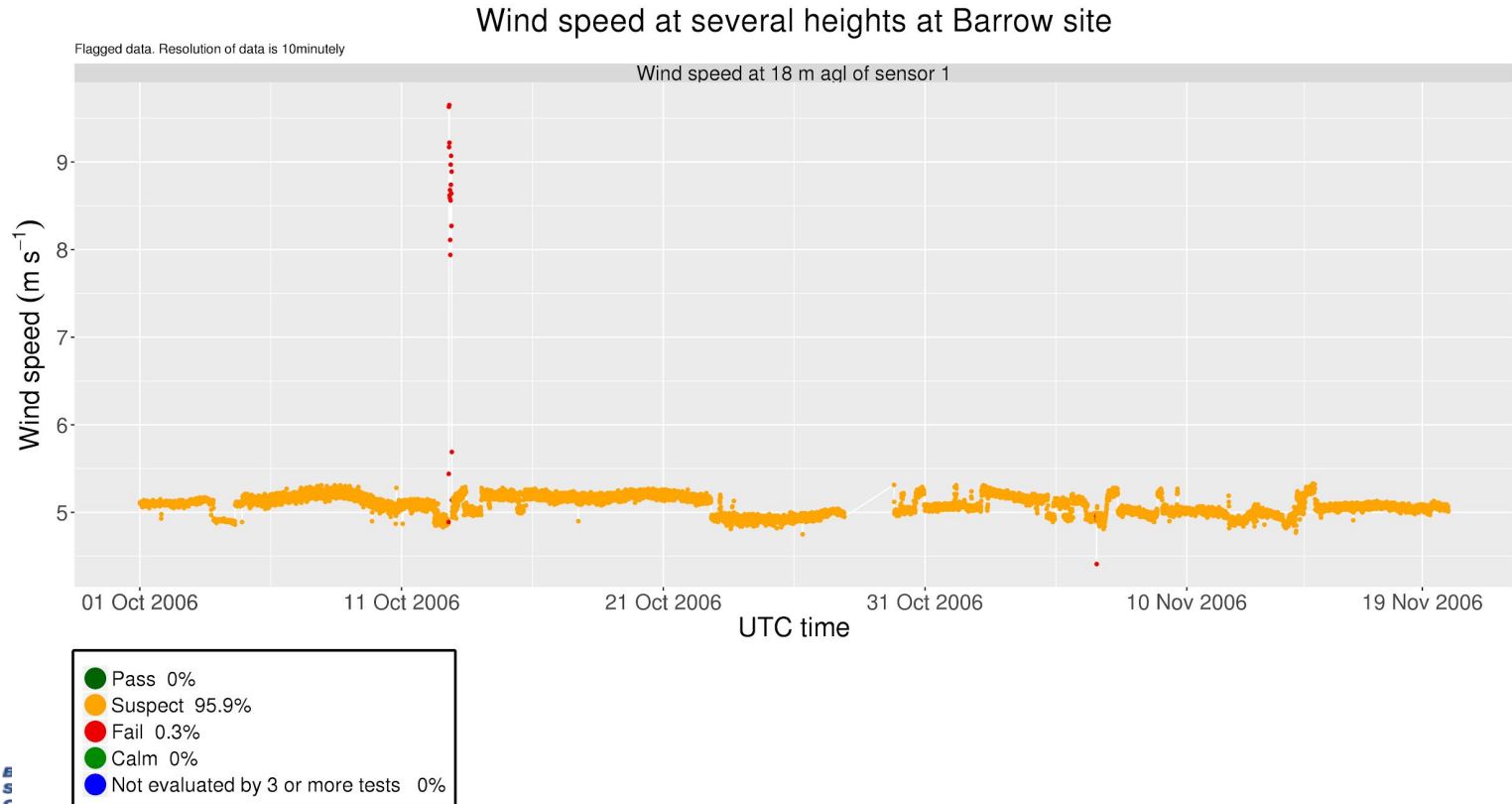
Flat lines

Wind speed at several heights at Butler Grade site



0% of data are missing

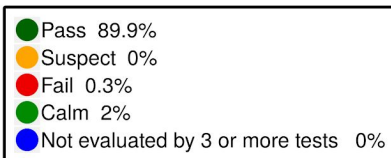
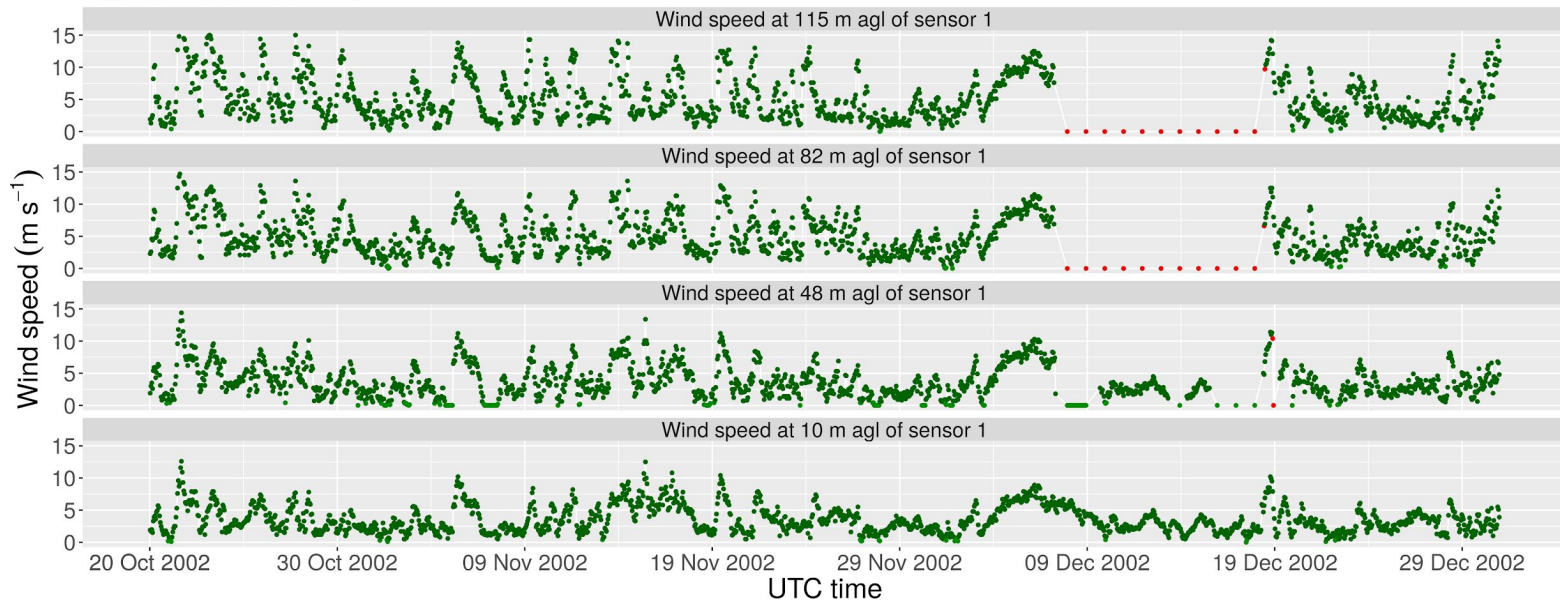
Spikes & abnormally low wind speed variability



Icing of the anemometer

Wind speed at several heights at Hegyhatsal site

Flagged data. Resolution of data is hourly



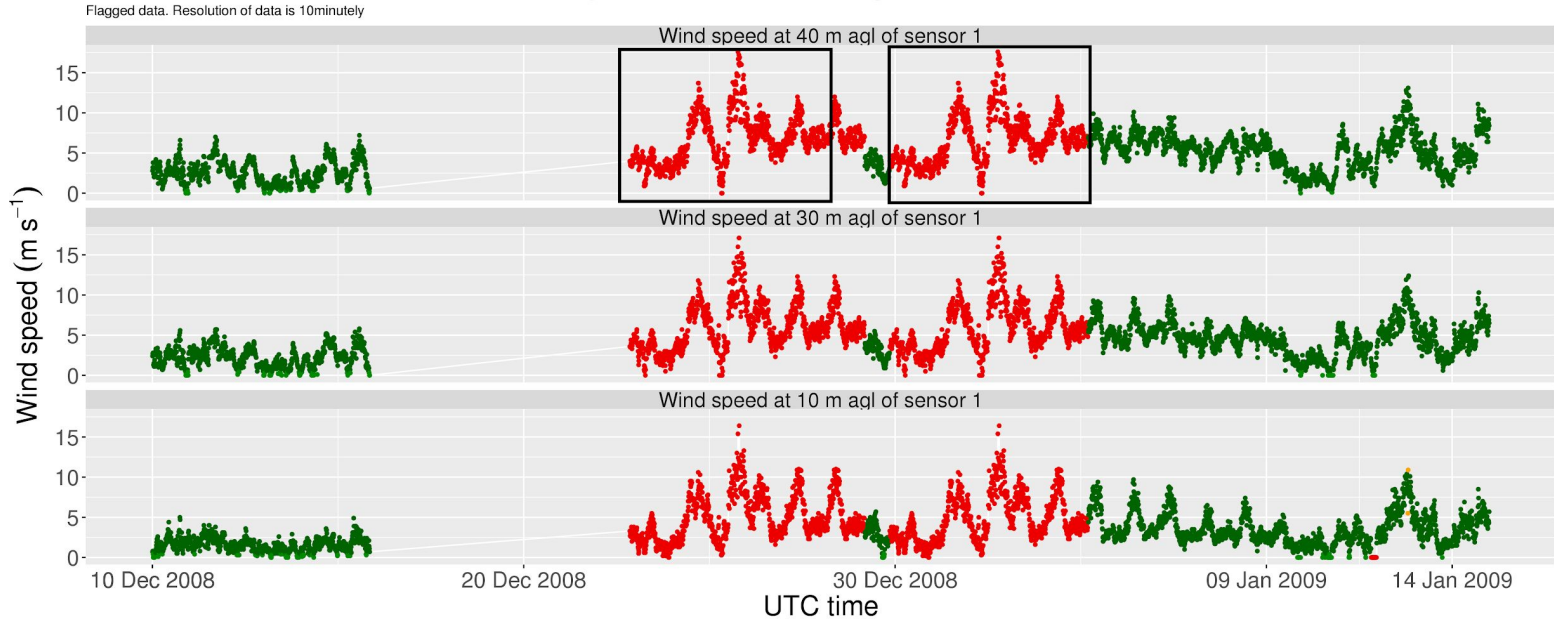
7.8% of data are missing



Duplicated sequences

a common but debatable procedure to fill in no-data periods

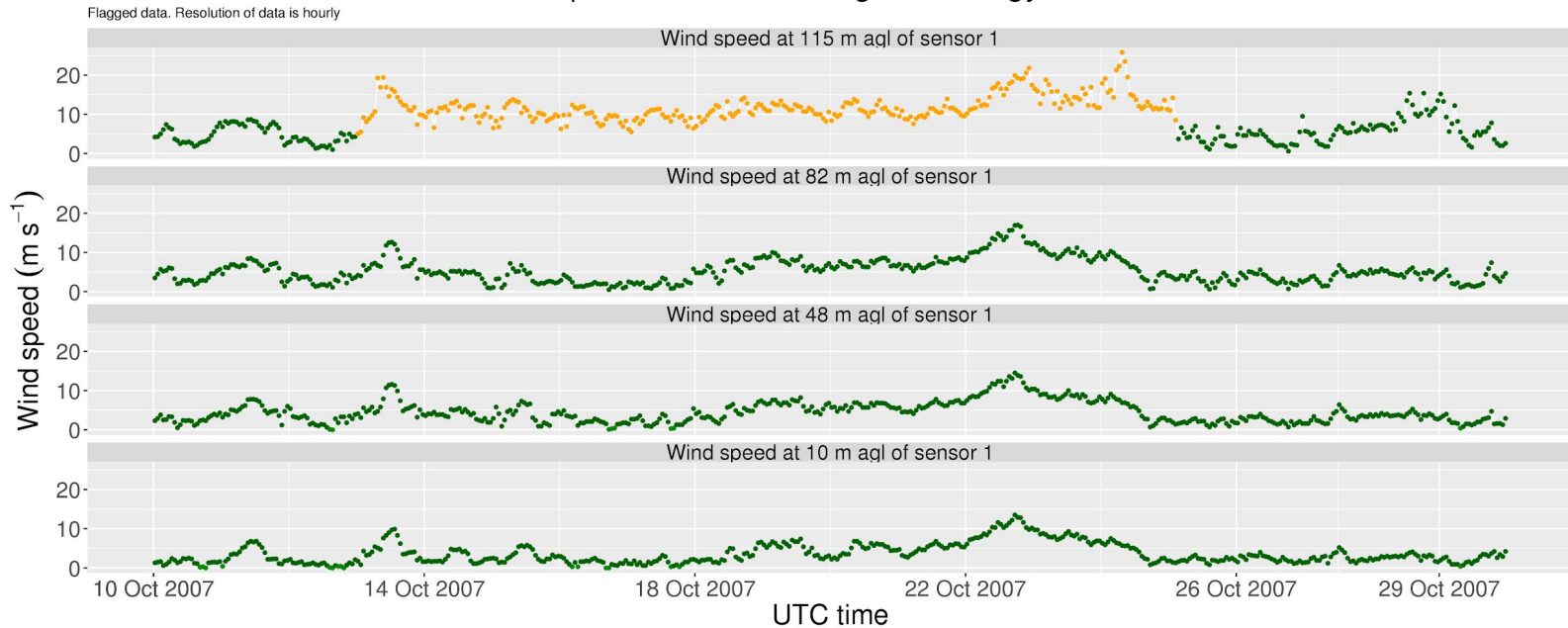
Wind speed at several heights at Abadan site



19.6% of data are missing

Systematic errors

Wind speed at several heights at Hegyhatsal site



0% of data are missing



The QC code should...

- 1** Look for abnormal statistical features in the series
- 2** Be conservative: remove **ONLY** the most gross errors
- 3** Be able to deal with high-frequency series



<https://essd.copernicus.org/articles/12/429/2020/essd-12-429-2020.html>

<https://earth.bsc.es/gitlab/jramon/INDECIS-QCSS4TT>

Q & A

Next meeting: 9th Apr. 2021 (Friday 3pm)