Using ML to predict regional renewable production

Gabriele Martinelli, Senior Analyst, European Power

Christopher Coello, Senior Analyst, European Power

joint work with Hilde Nyhus and Jørgen Hansen

7th Electricity Price Forecasting & Market Workshop, Jan 25, 2019

REFINITIV™

Agenda

- **Wind and solar power output modeling**
  - Current approach using actual weather data
  - Modelling using weather forecasts
  - Advanced methodologies (ML)

- **Dimensionality reduction and Feature selection**
  - Selection on spatial grid
  - Selection by modeling (sequential)
  - Selection by dimensionality reduction (PCA)

- **Results / case studies**
  - Wind power output in Germany
  - Wind power output in Nordics
  - Solar power output in Spain, PV and thermal and PV in Germany

- **Conclusions**

Motivation

Improved RES (renewable power output) models are crucial for better short term price forecasting (Intraday / DA / WA)

- **Current offering** from **Refinitiv** in Commodities → Power (Continental and Nordics):

1) Supply (hydro, wind, solar) and demand forecasts for (all) countries in Europe and price areas + AUS, US, Turkey, Brasil,...

2) 16+ weather runs per day (ECo and ECe, GFSo and GFSe + AROME / DWD / IKONeu,...). All data presented in EIKON and via feed. Historical data through Download Manager or PointConnect solutions

3) Availabilities and actual production figures (aggregated or per plant)

4) Actual prices and price forecasts for most countries in Nordics + CWE + ... using a European-wide fundamental model

5) DayAhead and WeekAhead analysis updated every day, longer term analysis (mid-term) updated once or twice a week.

6) Weather maps (temp, precip, wind,...) and comment from meteorologist twice a day.

7) Bid-offer curves and sensitivities

THOMSON REUTERS → REFINITIV

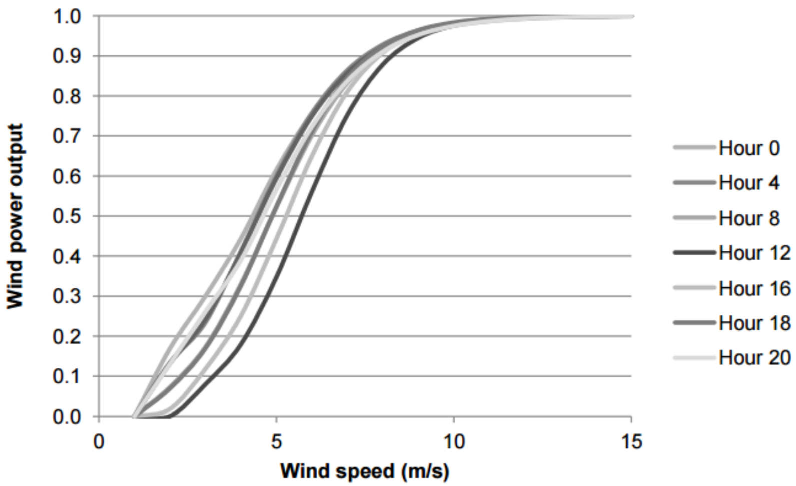
Wind and solar power output modeling

Current approach

- Physical model

$$\min_{\mathbf{w},\mu,\sigma} \sum_{h=01}^{24} \sum_{d=1}^{nd} \left(p_{h,d} - \sum_{st=1}^{nst} \left(w_{st} \cdot \frac{1}{1+e^{-\frac{x_{h,d,st}-\mu_{h,st}}{\sigma_{h,st}}}} \right) \right)^2$$

where:
- ω are the weights associated to each station
- p is the wind power production
- x is the wind speed
- μ and σ are the parameters associated to the sigmoid function (per hour)



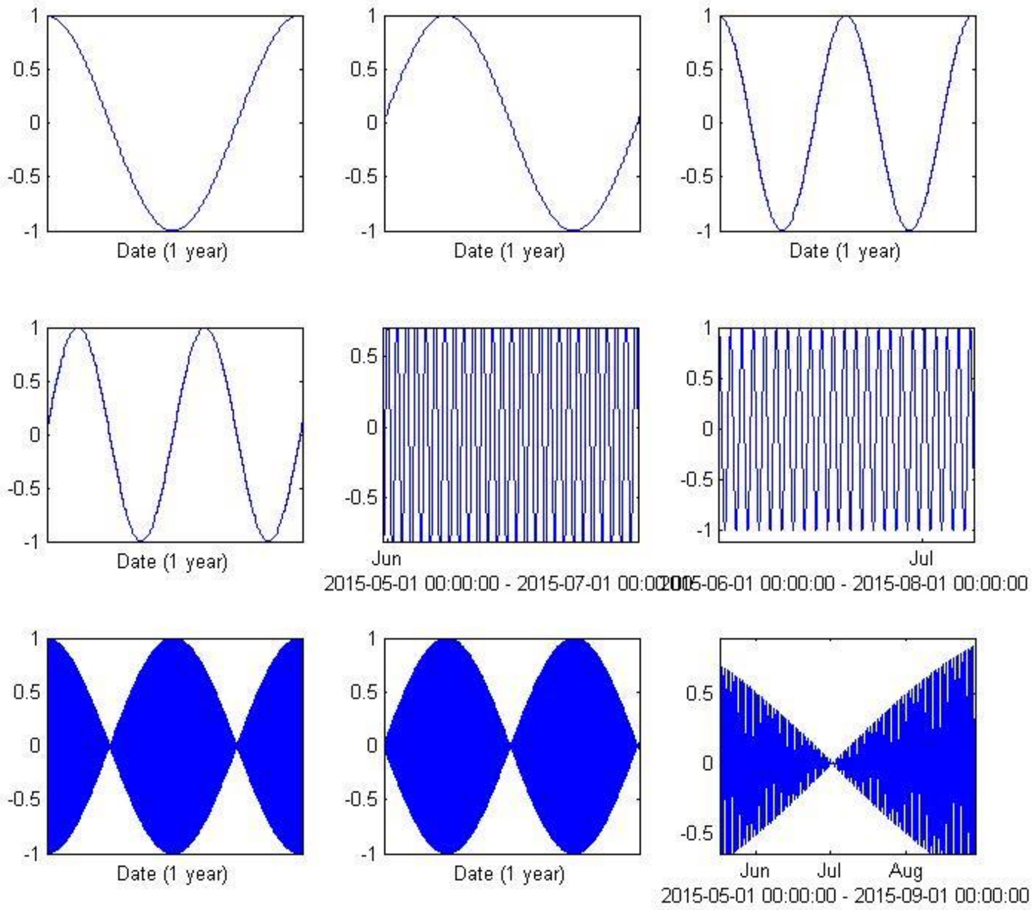
Current approach

- Physical model with Fourier seasonality correction

$$\min_{\mathbf{w_{f,st}}} \sum_{d=1}^{nd}\sum_{h=01}^{24}\left(lf_{h,d} - \sum_{f=1}^{nf}\sum_{st=1}^{nst} w_{f,st}\boxed{f_{f,st}}\boxed{\frac{1}{1+e^{-\frac{x_{h,d,st}-\mu}{\sigma}}}}\right)^2$$

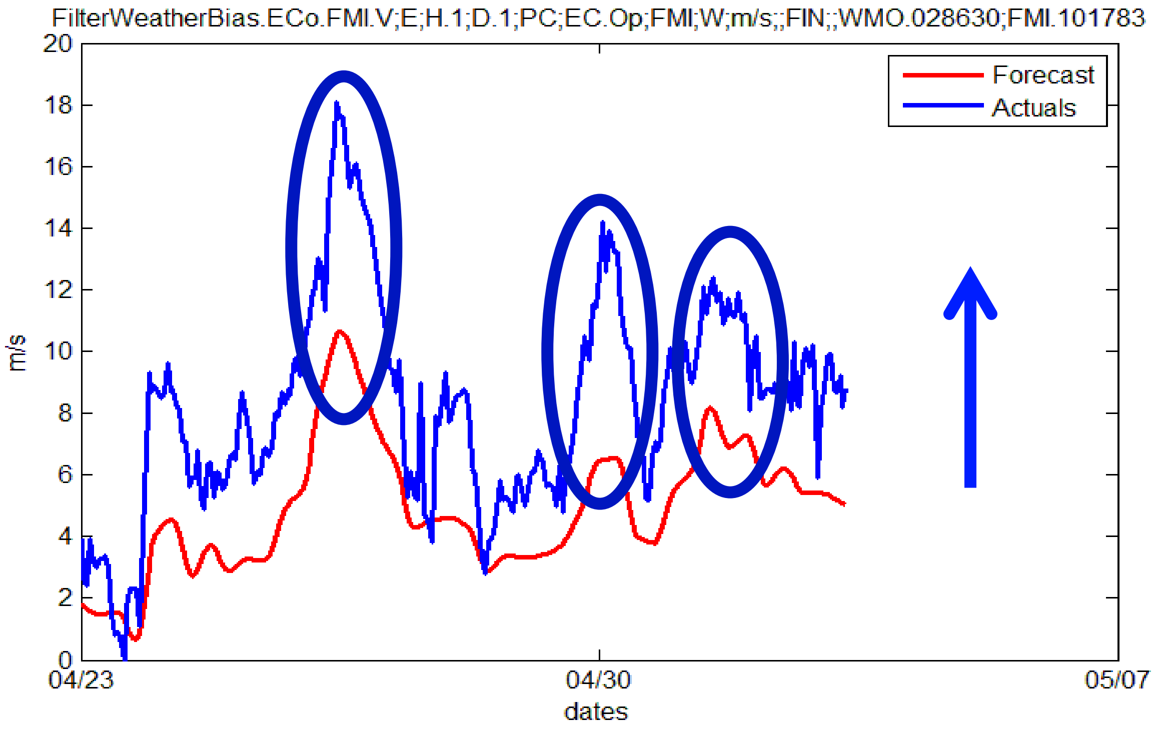
where:

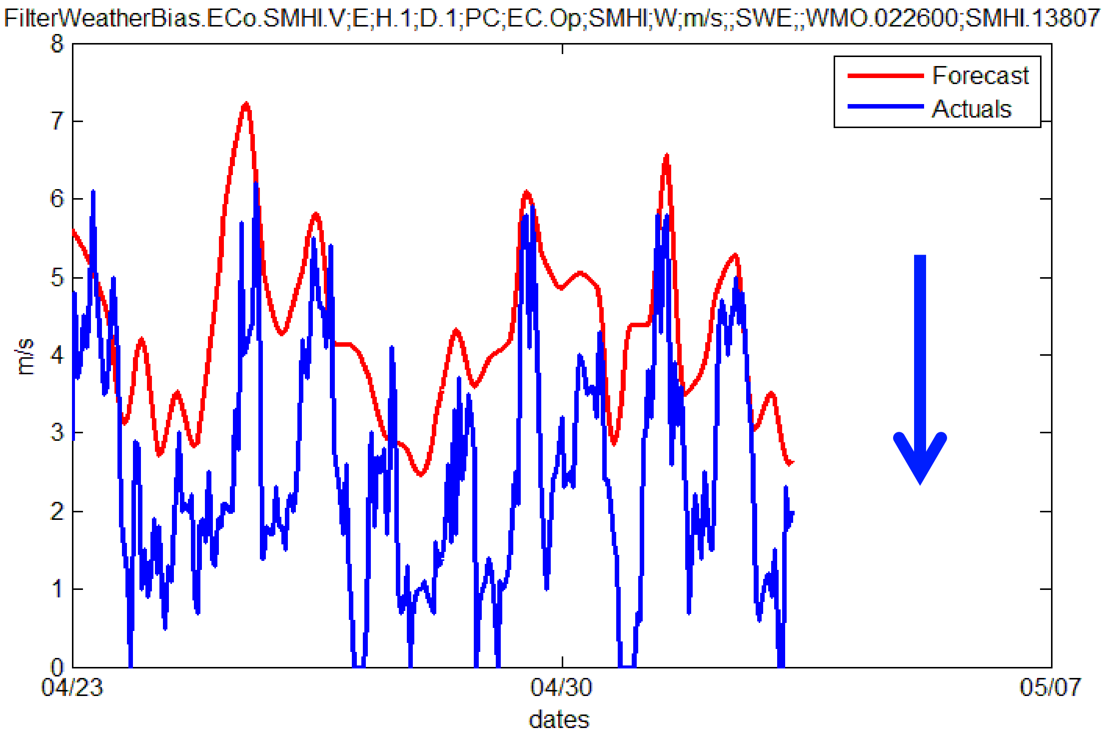
- ω are the weights associated to each station and fourier basis
- lf is the load factor
- x is the wind speed
- μ and σ are the parameters associated to the sigmoid function (fixed, to ensure the linearity of the problem)
- f are the Fourier basis (see right picture) used for projection



Current approach

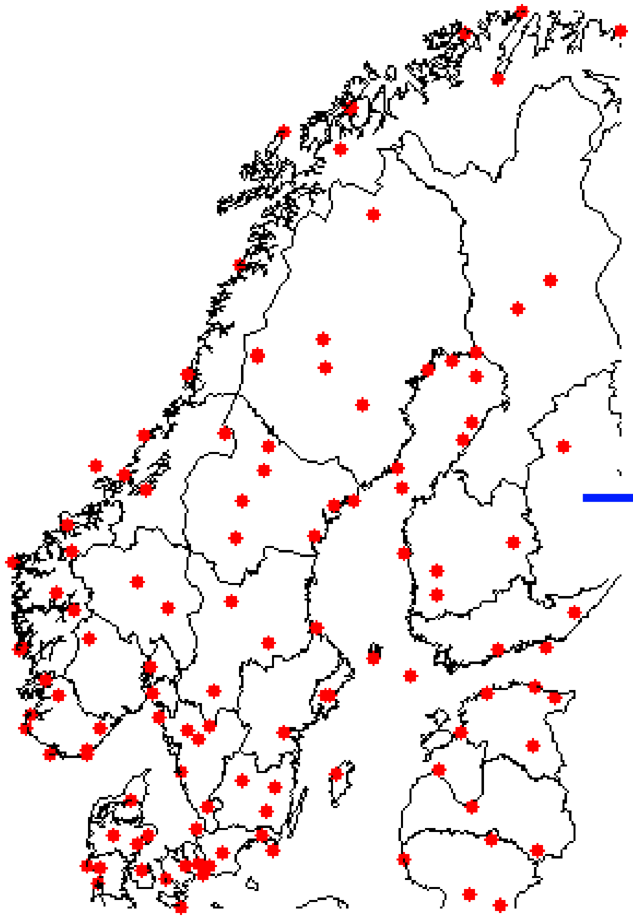
- Problem with actual vs forecast data → Filtering

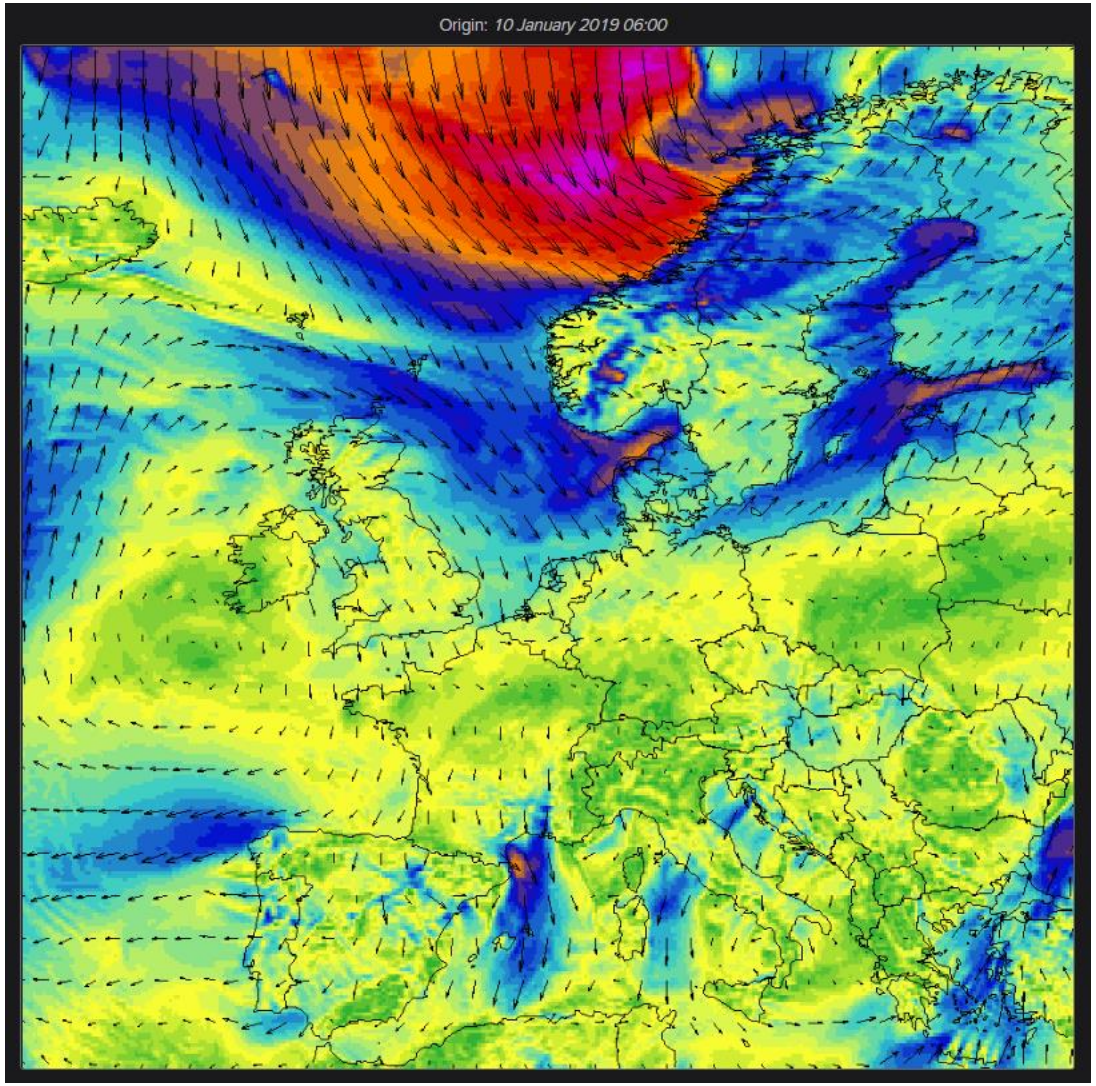




Modeling with weather forecast

Actual weather stations





Size ECo EU grid:
- ~ 1 grid point / 7km (EW)
- ~ 1 grid point / 14km (NS)

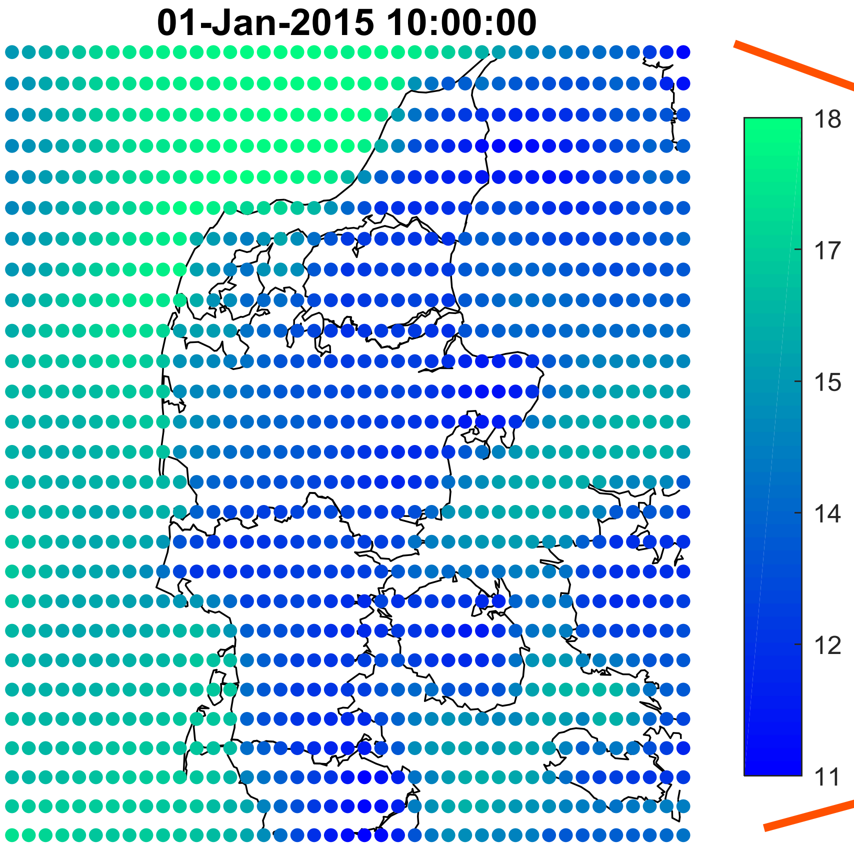
→ 1000 grid points for DK1
→ 4200 points for FIN
→ 4600 points for DEU

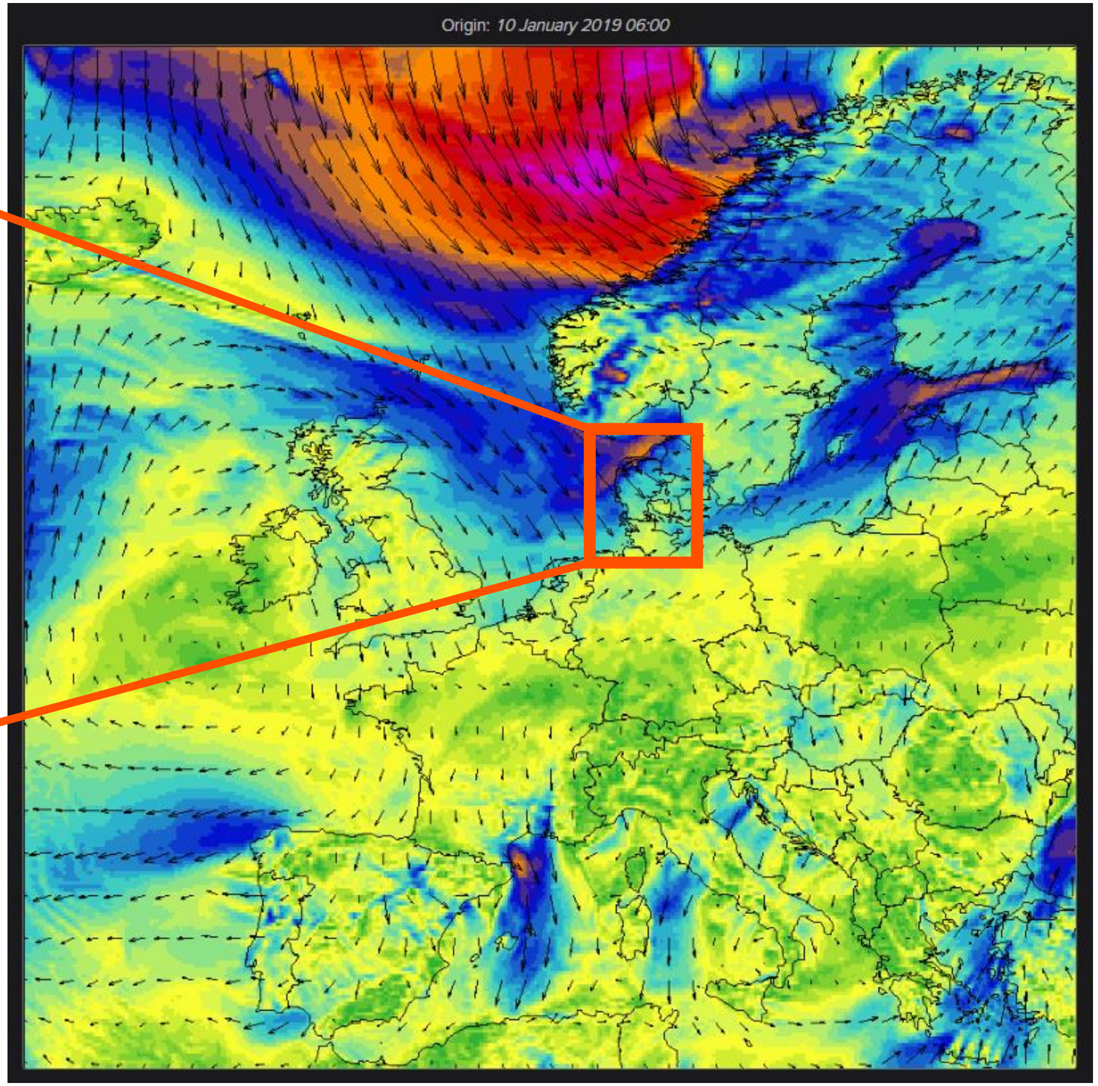
1 point per hour →
100M points per variable (FIN)

→ **~1B points** per country assuming 10 weather variables per calibration run.

REFINITIV

Modeling with weather forecast

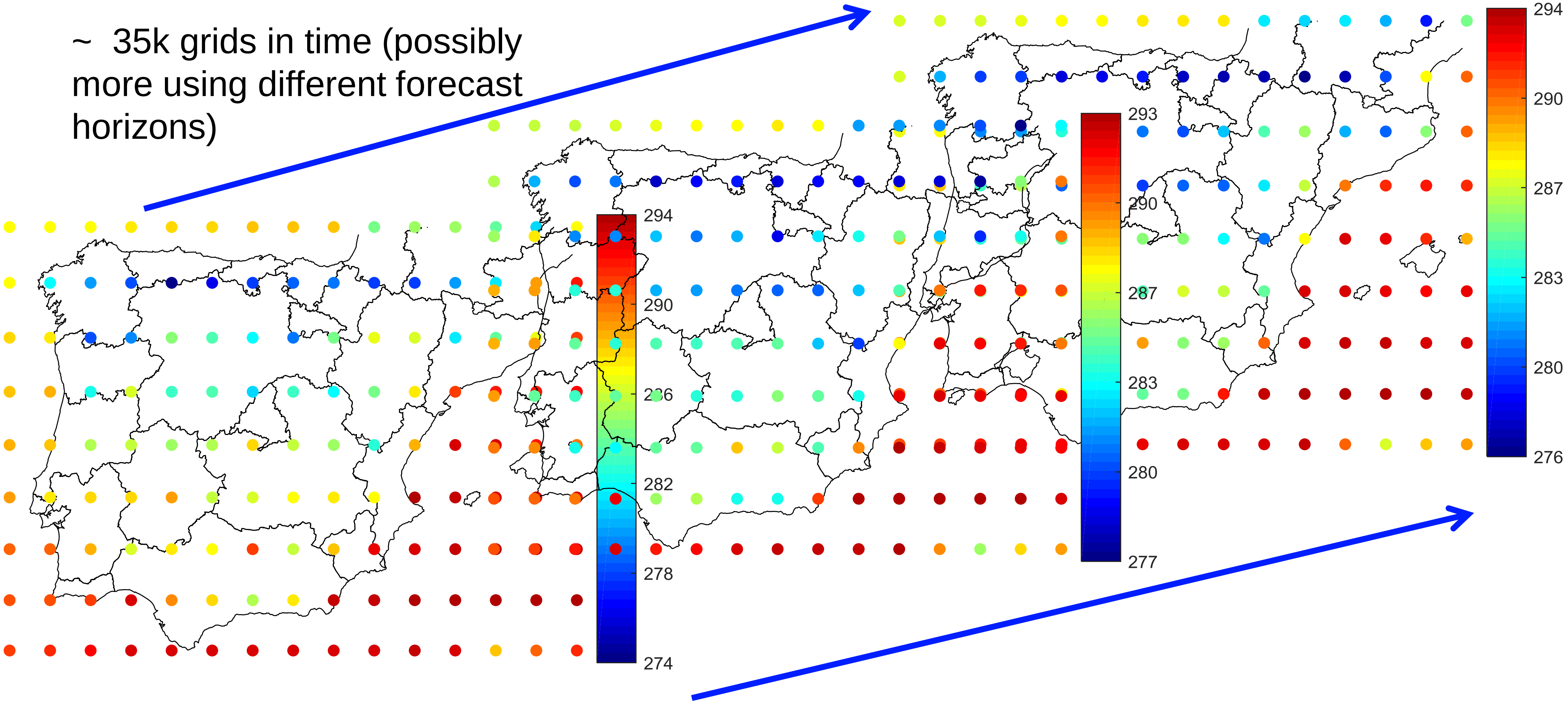




~1B points per country → ML methods for reducing the dimension of the problem and extract max info from the grid

Modeling with weather forecast

~ 35k grids in time (possibly more using different forecast horizons)

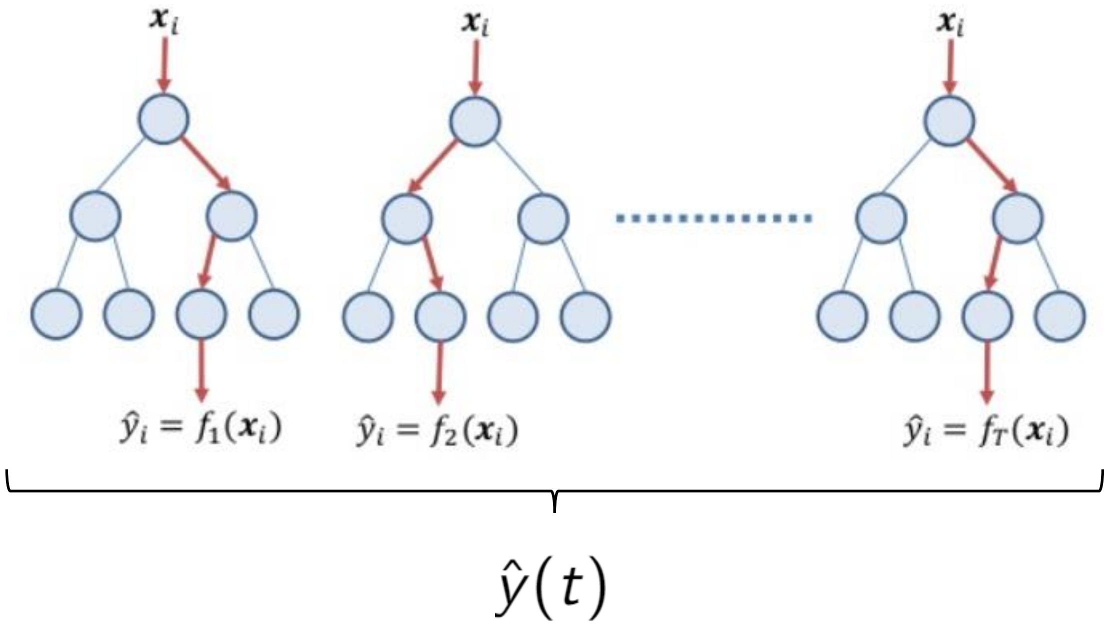


REFINITIV

Random forests for wind power output estimation

Theory

An ensemble of decision trees trained by bootstrap sampling and random feature selection



Implementation

MATLAB

Machine Learning and Statistics Toolbox

Feature selection → ~10 minutes with sequential algorithm

Training with optimisation → 1 to 2 hours

Prediction → ms to s

Optimization hyperparamters

- Minimum number of points per leaf
- Number of decision features per split
- Maximum number of splits

REFINITIV

Random forests for wind power output estimation

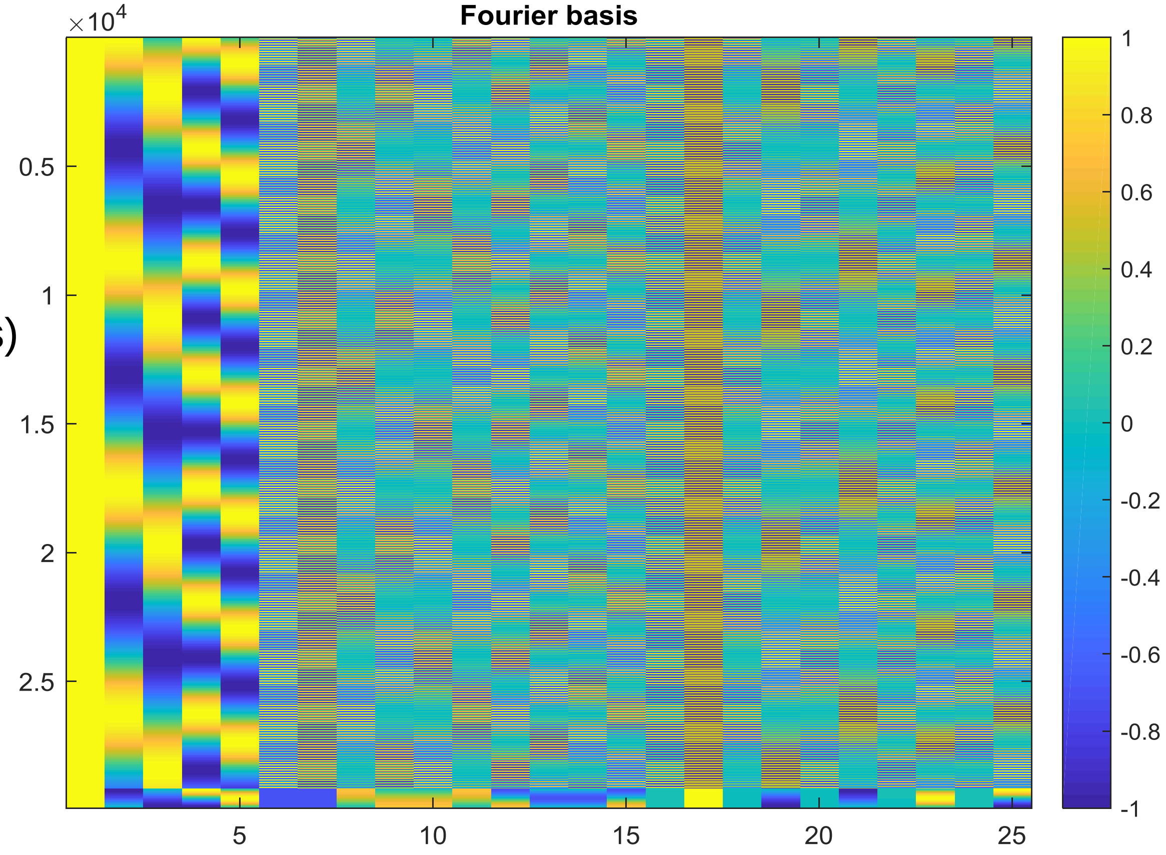
INPUTS

Forecast : ECoEU (denser grid) + date variables (Fourier)

Variables : 100m U and V components (possibly Temp)

Resolution : hourly data

Model output : 00 and 12 (plus 06z and 18z for the last 3 months)



Random forests for wind power output estimation

INPUTS

Forecast : ECoEU (denser grid) + date variables (Fourier)

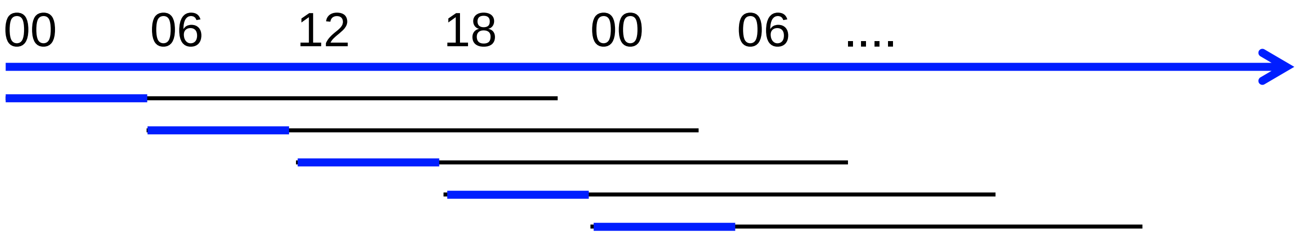
Variables : 100m U and V components (possibly Temp)

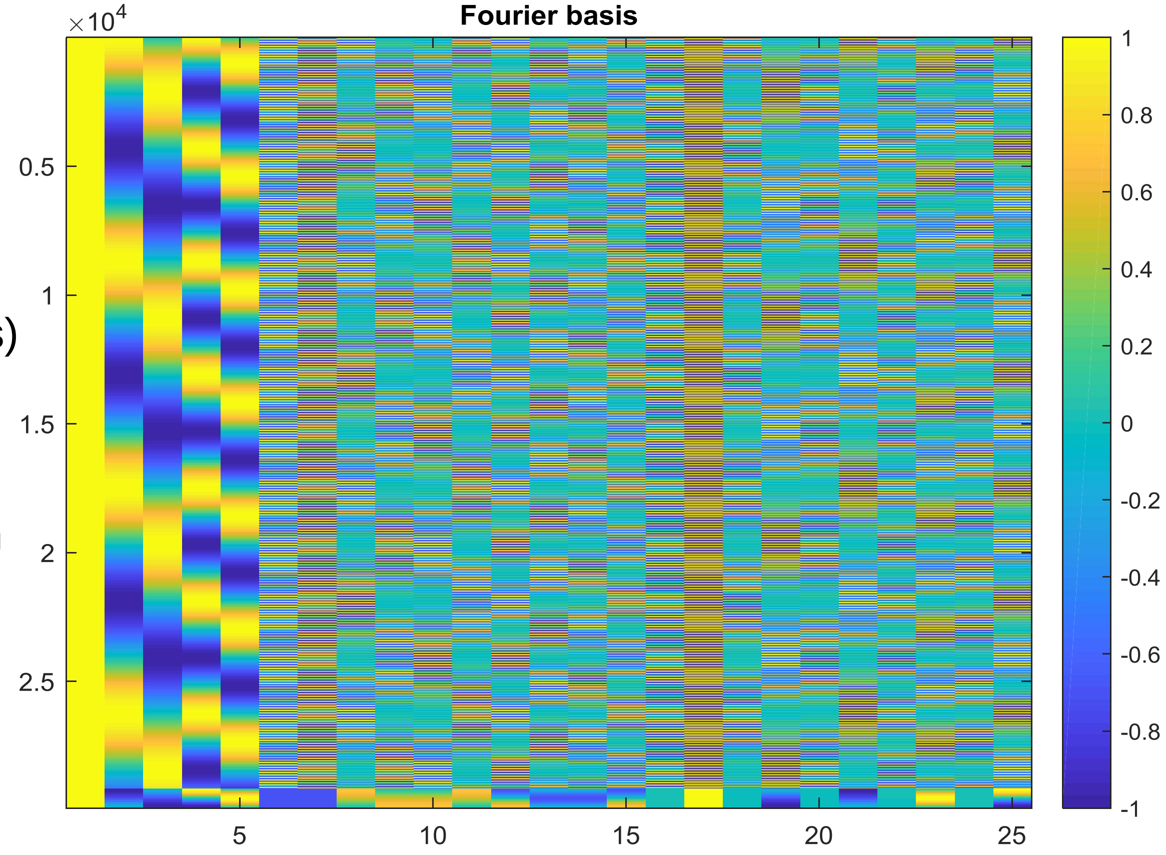
Resolution : hourly data

Model output : 00 and 12 (plus 06z and 18z fo rthe last 3 months)

Pre-processing:

- Assemble a timeserie by concatenating the first 12/6 hours of each model output





Random forests for wind power output estimation

INPUTS

Forecast : ECoEU (denser grid) + date variables (Fourier)

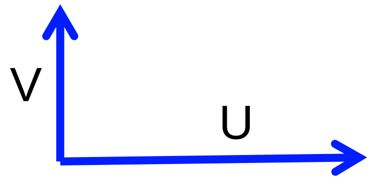
Variables : 100m U and V components (possibly Temp)

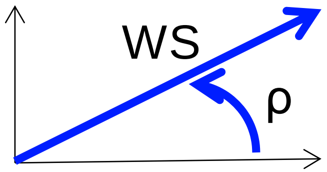
Resolution : hourly data

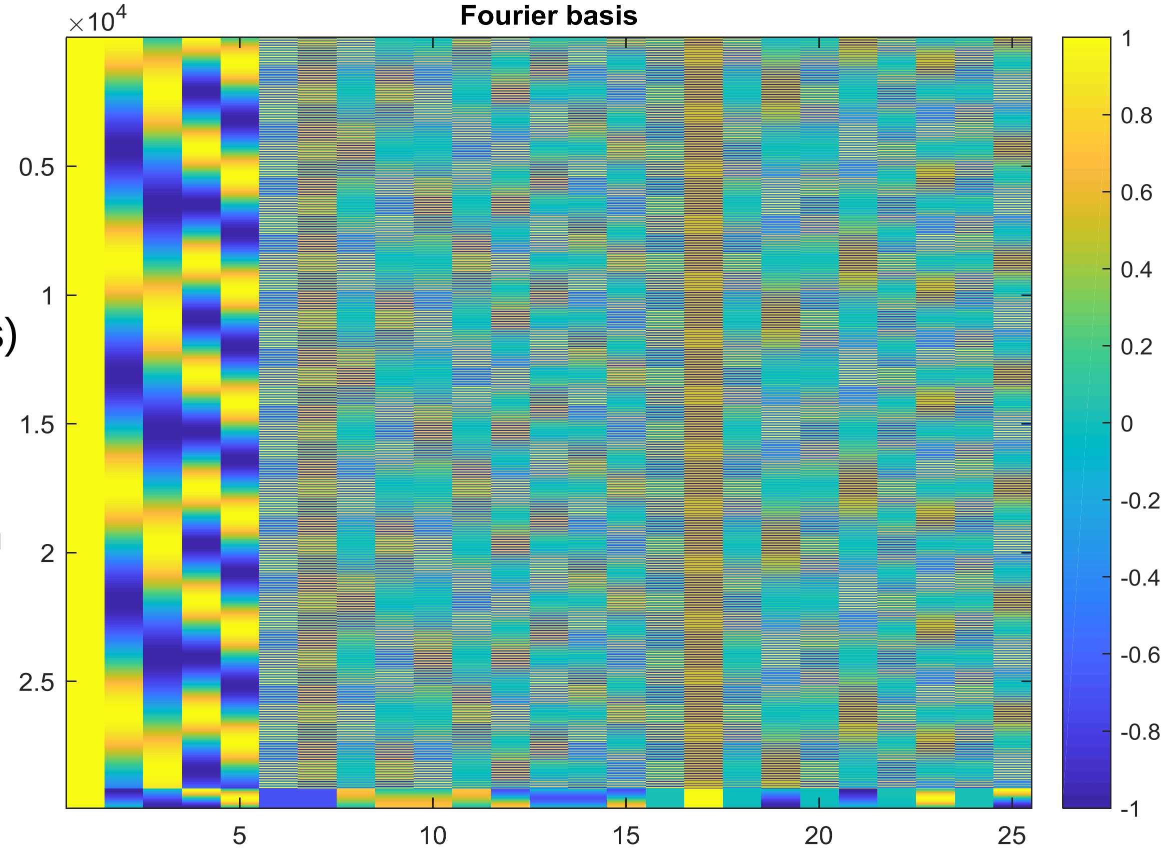
Model output : 00 and 12 (plus 06z and 18z for the last 3 months)

Pre-processing:

- Assemble a timeserie by concatenating the first 12/6 hours of each model output
- Change the base of the representation from the (U,V) components to the polar coordinates (WS,ρ)







REFINITIV

Random forests for wind power output estimation

INPUTS

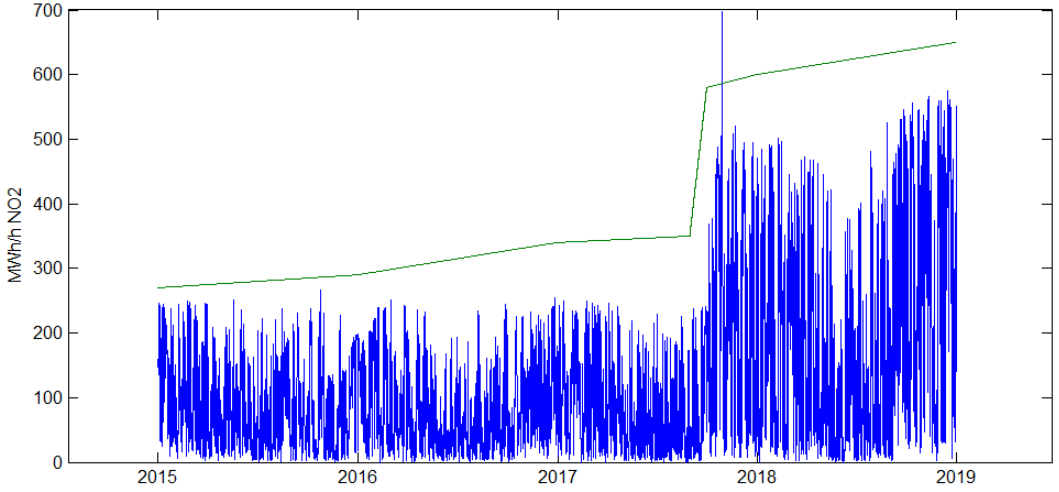
Forecast : ECoEU (denser grid) + date variables (Fourier)
Variables : 100m U and V components (possibly Temp)
Resolution : hourly data
Model output : 00 and 12 (plus 06z and 18z for the last 3 months)

Pre-processing:

- Assemble a timeserie by concatenating the first 12/6 hours of each model output
- Change the base of the representation from the (U,V) components to the polar coordinates (WS,ρ)

OUTPUTS

Load factor : Hourly production divided by the installed capacity



REFINITIV

Solar power output modeling

$$SP(t) = Cap(t) * \sum_{st=1}^{nst} \omega_{st} * f(\alpha_{st}, \theta_{st}, SR_{st}, t) * \eta_{st}(T_{st}, t)$$

where:

- SP(t) = solar production at time t
- ω are the weights associated to each station/location
- Cap(t) is the capacity of the region at time t
- SR is the solar radiation in W per square meter
- η is the efficiency of the location, which is a function of the solar farm design and of the temperature at time t

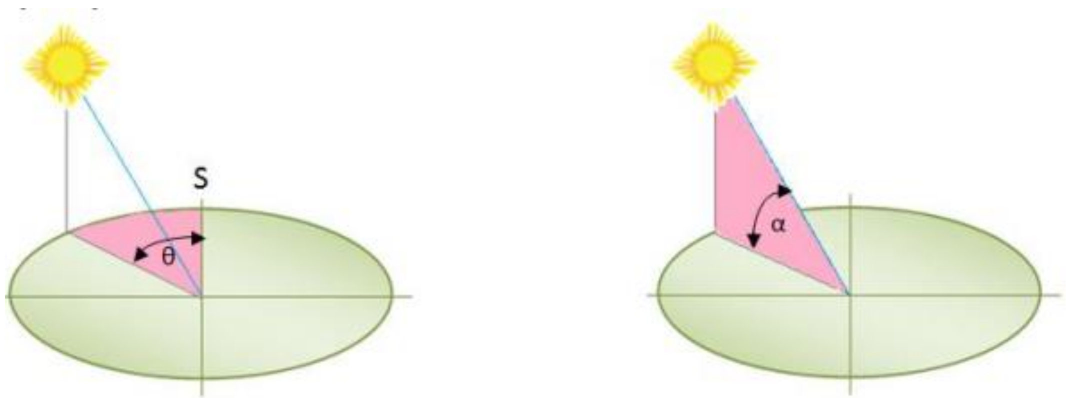


Figure 1: Azimuth and solar elevation angles definition

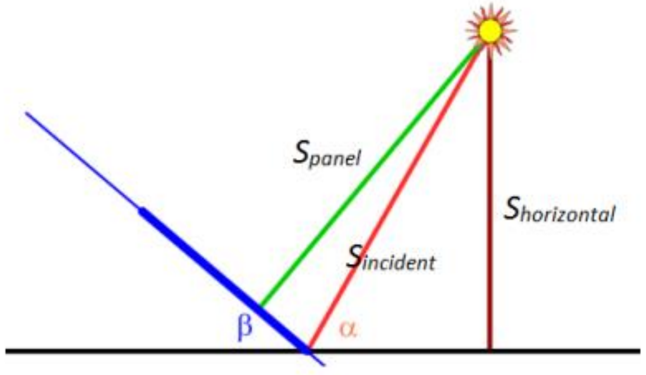


Figure 2: Horizontal, perpendicular and incident radiation

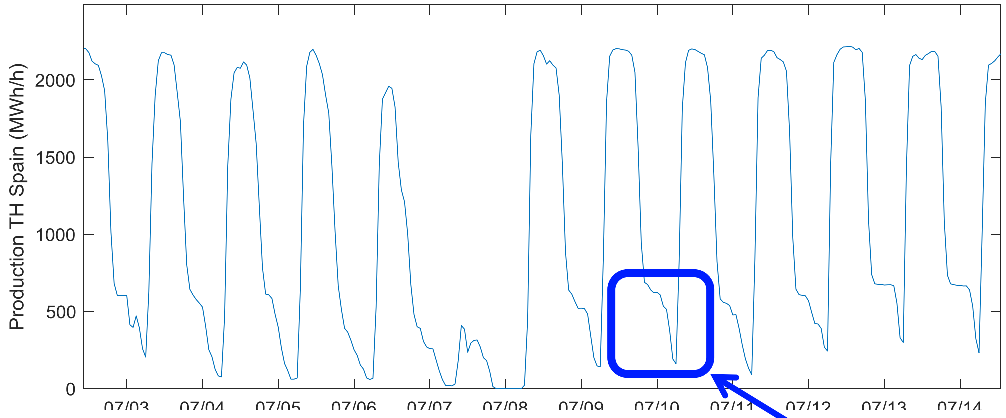
REFINITIV

Solar power output modeling: limitations for thermal component

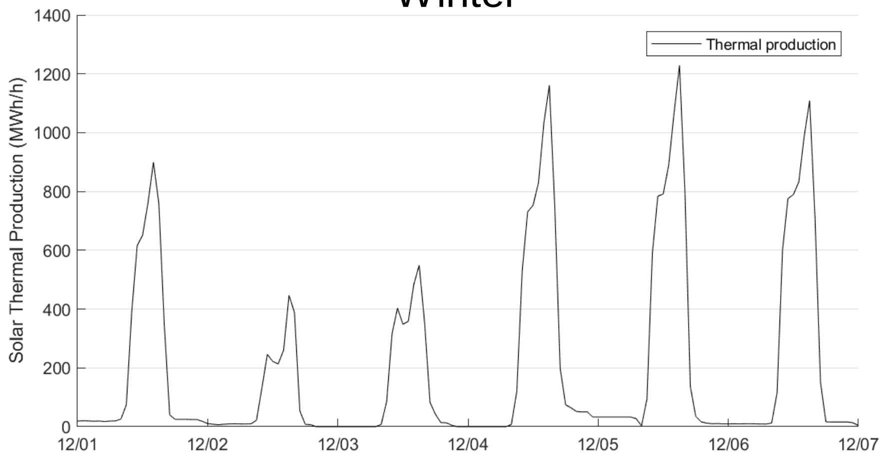
$$SP(t) = Cap(t) * \sum_{st=1}^{nst} \omega_{st} * f(\alpha_{st}, \theta_{st}, SR_{st}, t) * \eta_{st}(T_{st}, t)$$

- *f* is based on theoretical shape that forces zero production at night.
- No easy extension to Spanish thermal production: output during night in summer, but not in winter.

Summer



Winter



Night production

REFINITIV

Solar power output modeling with random forests

INPUTS

Forecast : GFSo

Variables :

- Solar radiation
- Temperature
- Cloud cover
- ”Theoretical shape” (to assign sunrise and sunset)

Resolution : 3h data

Model output : 00 , 06, 12 and 18.

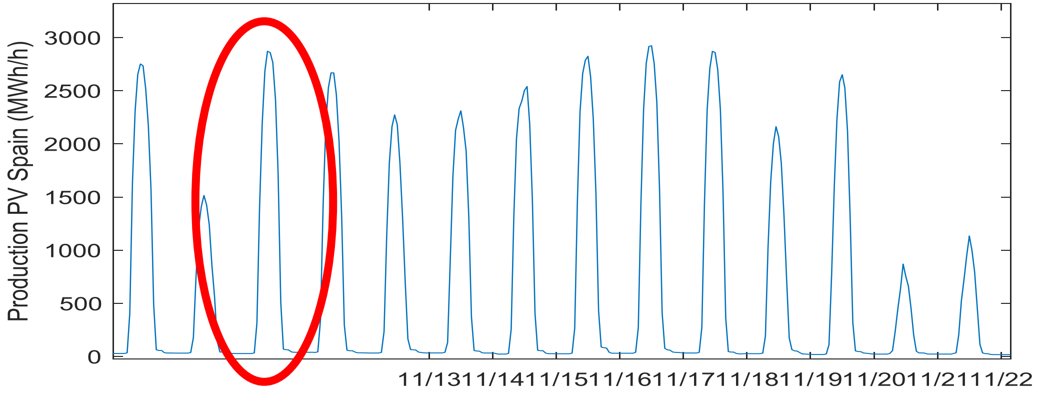
Pre-processing:

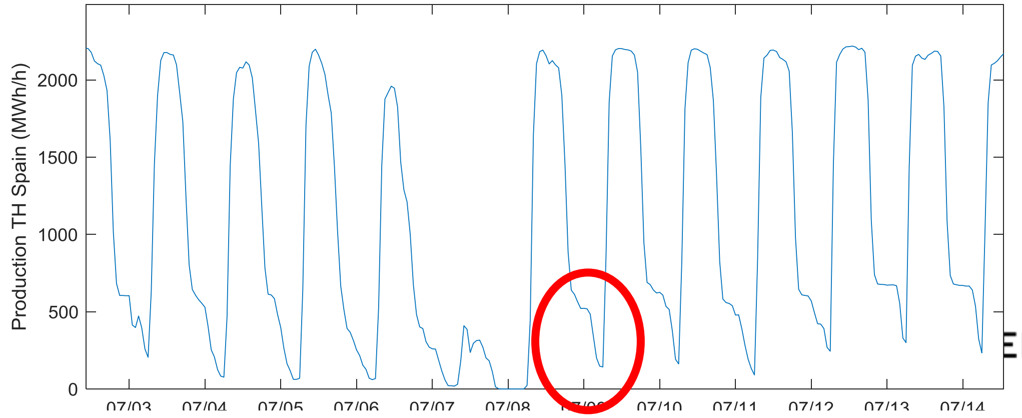
Assemble a timeserie by concatenating the first 12 hours of each model output

OUTPUTS

Load factor :

Hourly production divided by the installed capacity

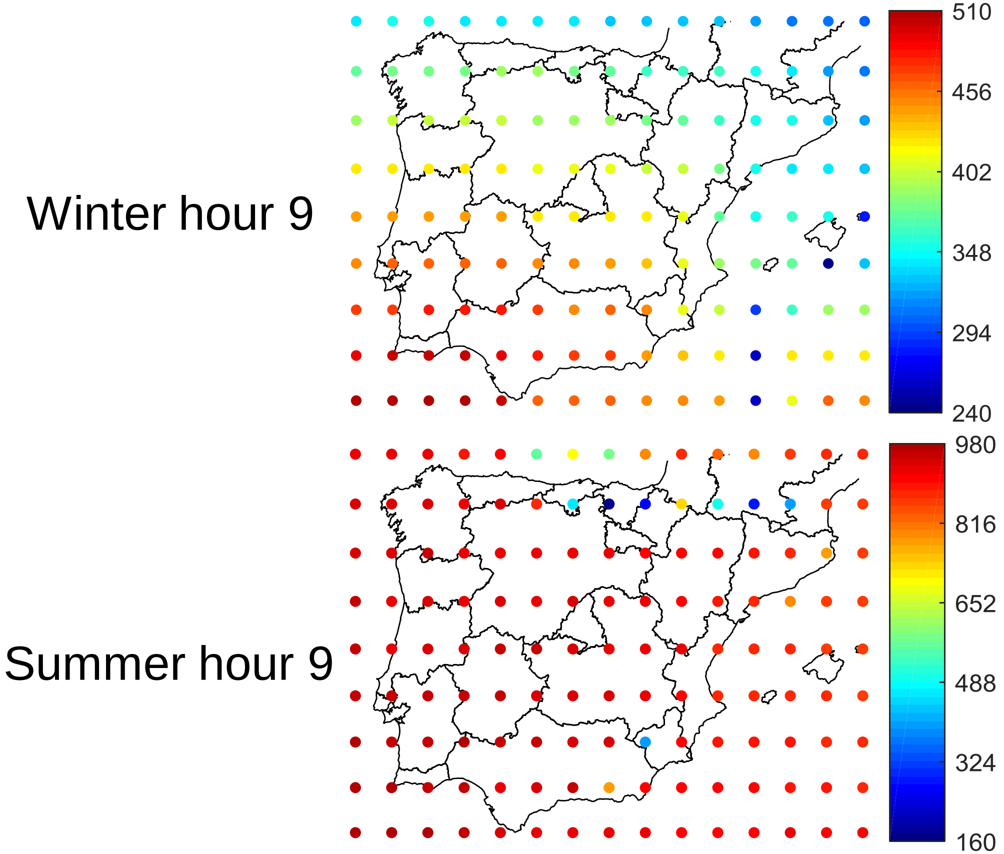


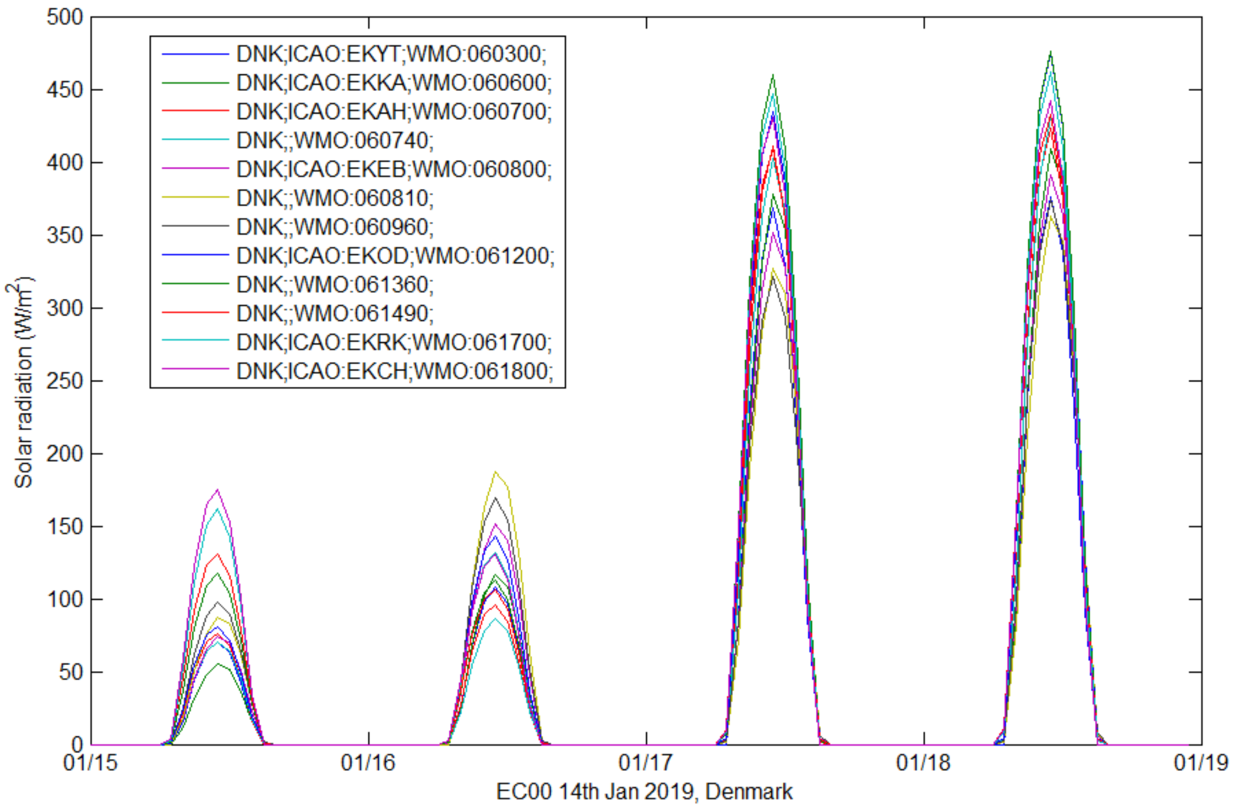


Solar power output modeling

INPUTS

Solar radiation



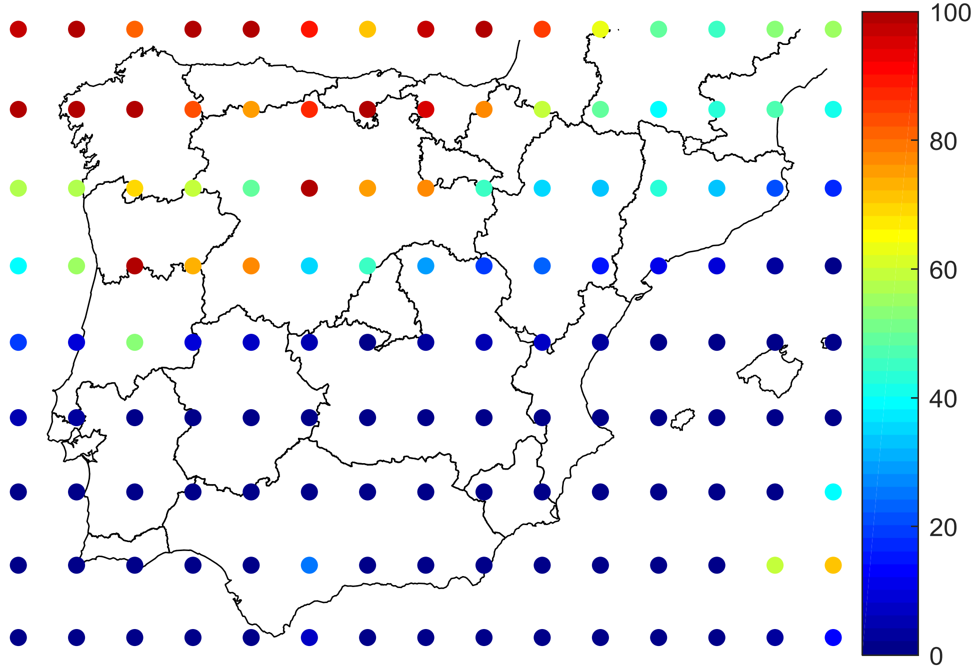


REFINITIV

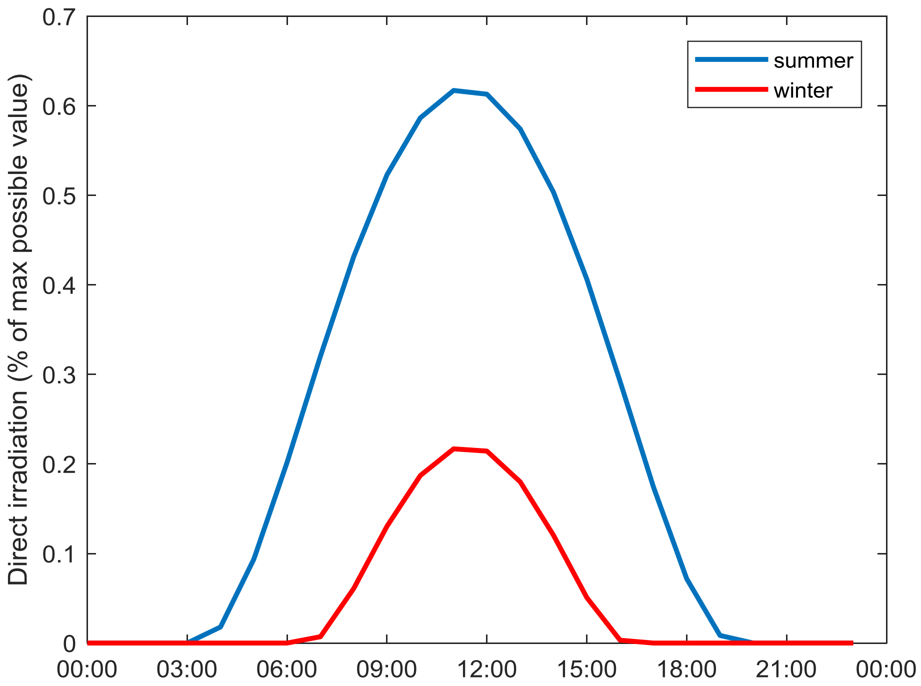
Solar power output modeling with random forests

INPUTS

Cloud cover



Theoretical shape

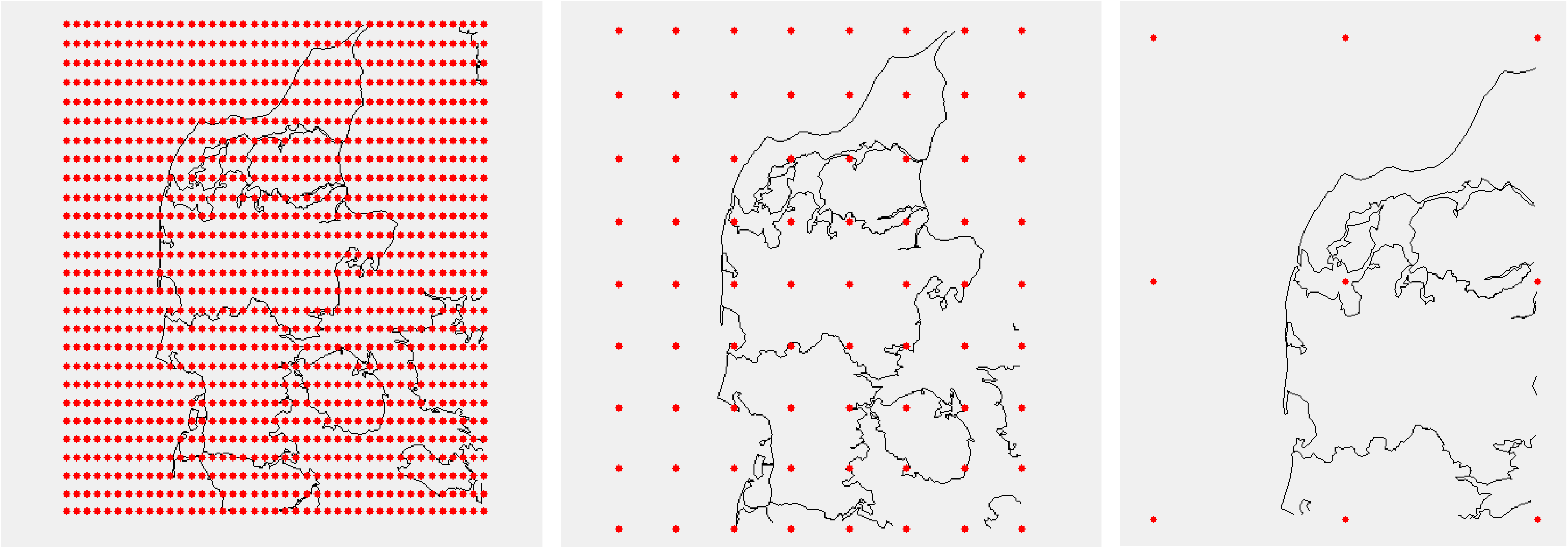


Dimensionality reduction and feature selection

REFINITIV™

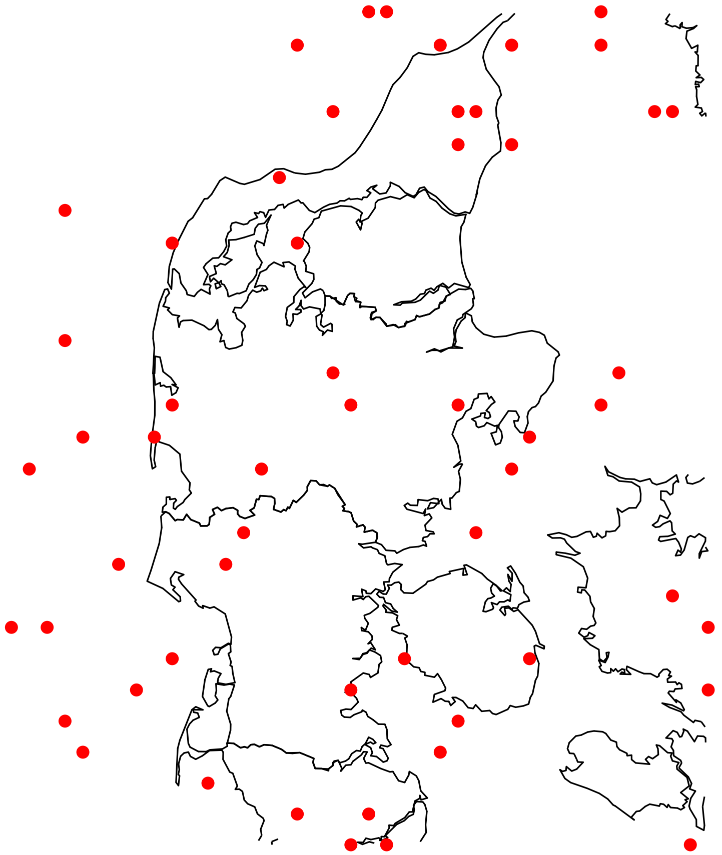
Selection on the spatial grid

- Grid subsampling



Selection on the spatial grid

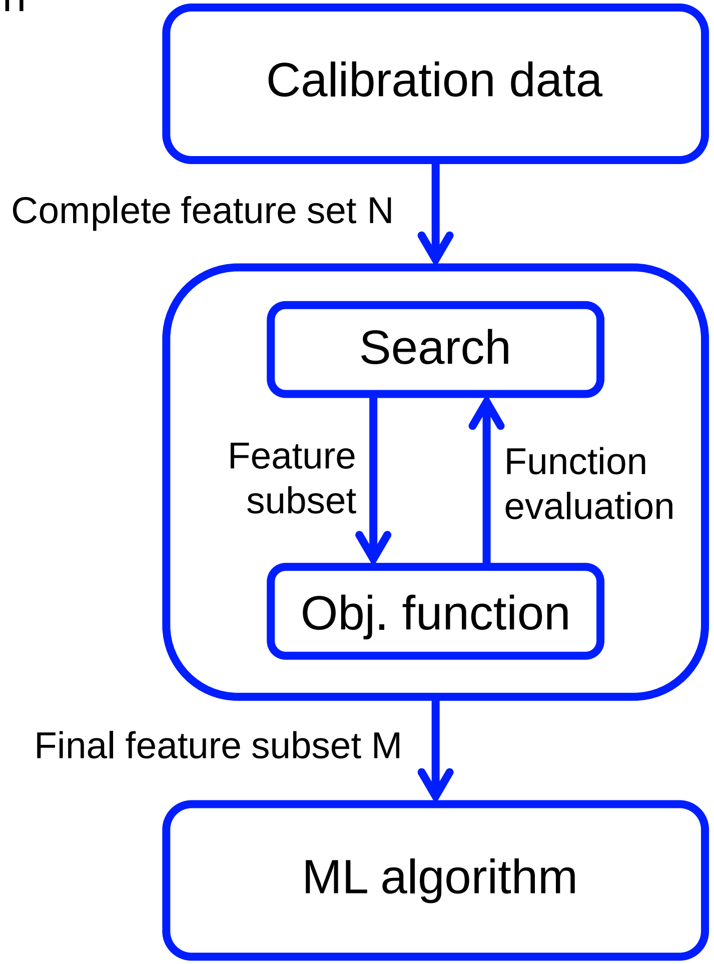
- Randomized



Selection by modelling

- **Idea**: Perform a ”**pre-calibration**” with a simplified function in a sequential (or joint) way in order to determine the best features

- **Selection by modelling requires:**
  – A search strategy to select candidate subsets
  – An objective function to evaluate these candidates



Selection by modeling

- **Idea**: Perform a ”**pre-calibration**” with a simplified function in a sequential (or joint) way in order to determine the best features

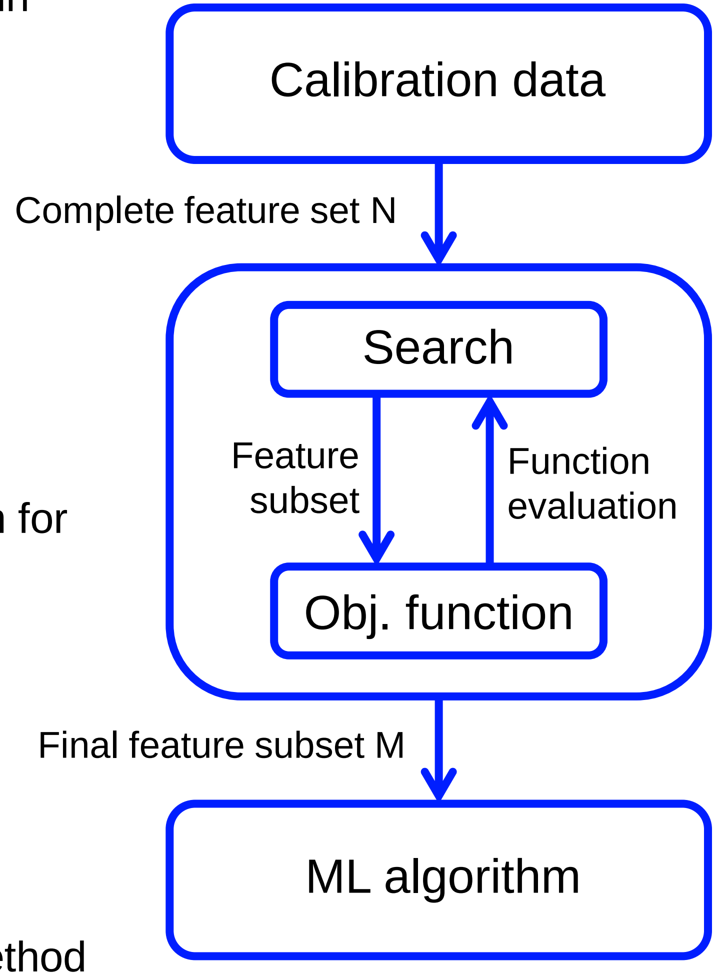
- **Selection by modelling requires:**
  – A search strategy to select candidate subsets
  – An objective function to evaluate these candidates

- **Search strategy**

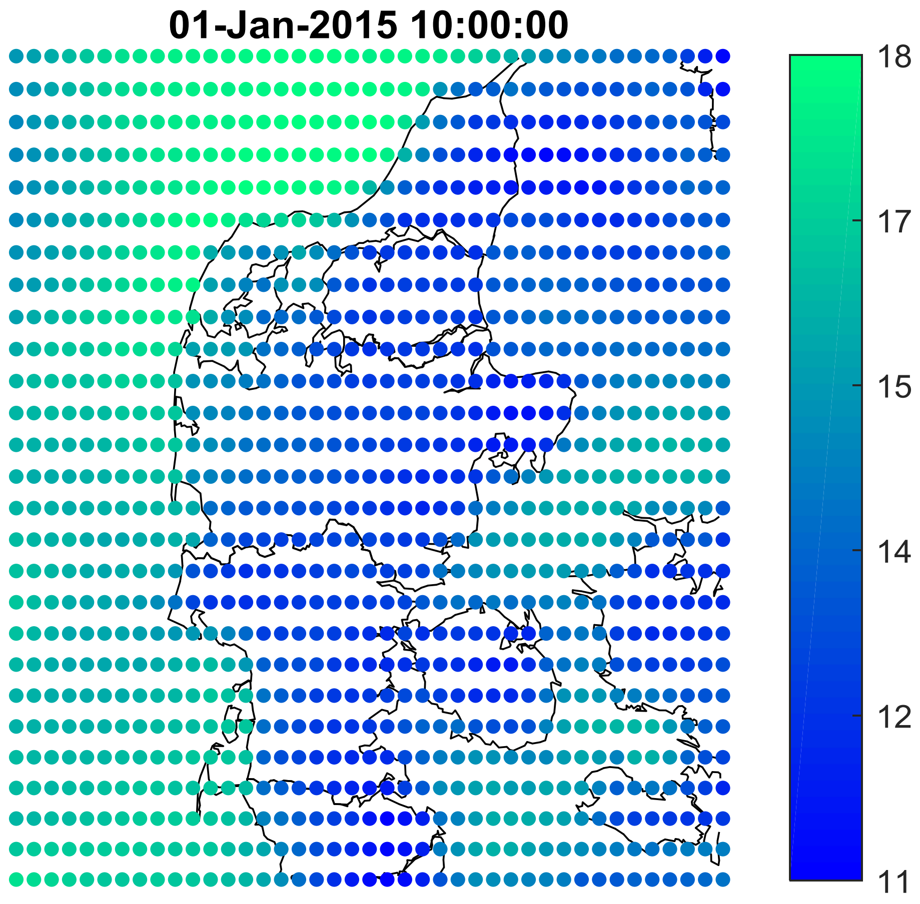
– Exhaustive evaluation of feature subsets involves an unfeasible # of combinations (even for moderate # of N total features and M features to select), so a search procedure is needed

– We have chosen a *sequential forward strategy* (several other choices are possible…)

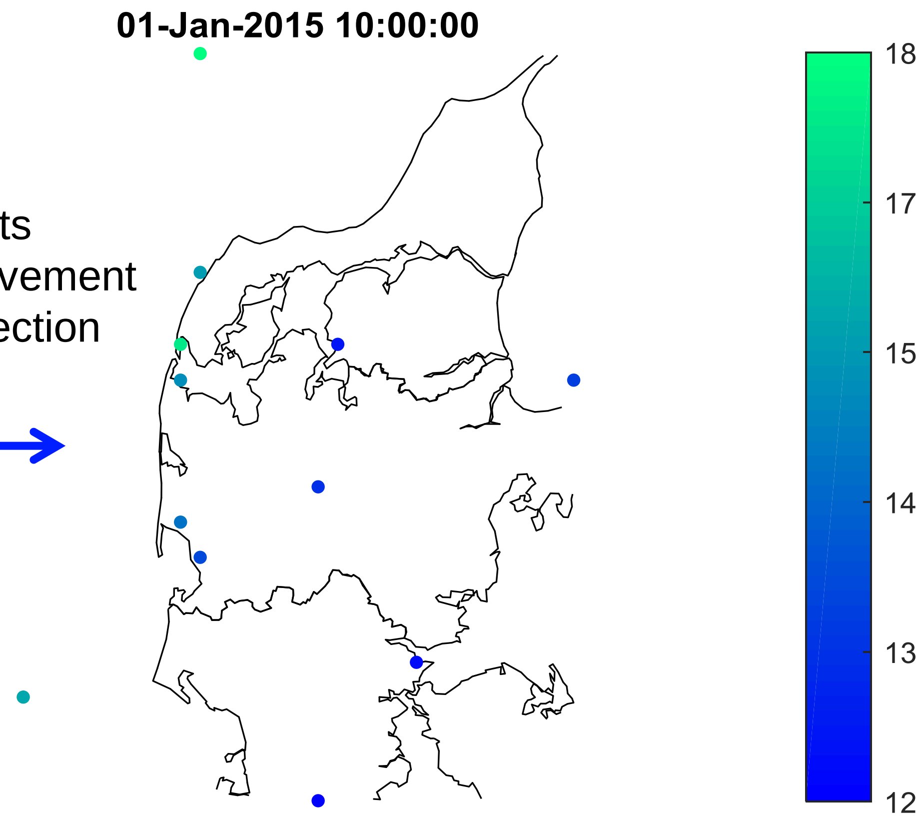
- **Objective function**
  – The objective function evaluates candidate subsets and returns a measure of their “goodness”, a feedback signal used by the search strategy to select new candidates.
  – A simplified version of random forest can be used, otherwise some clever regression method



Selection by modelling

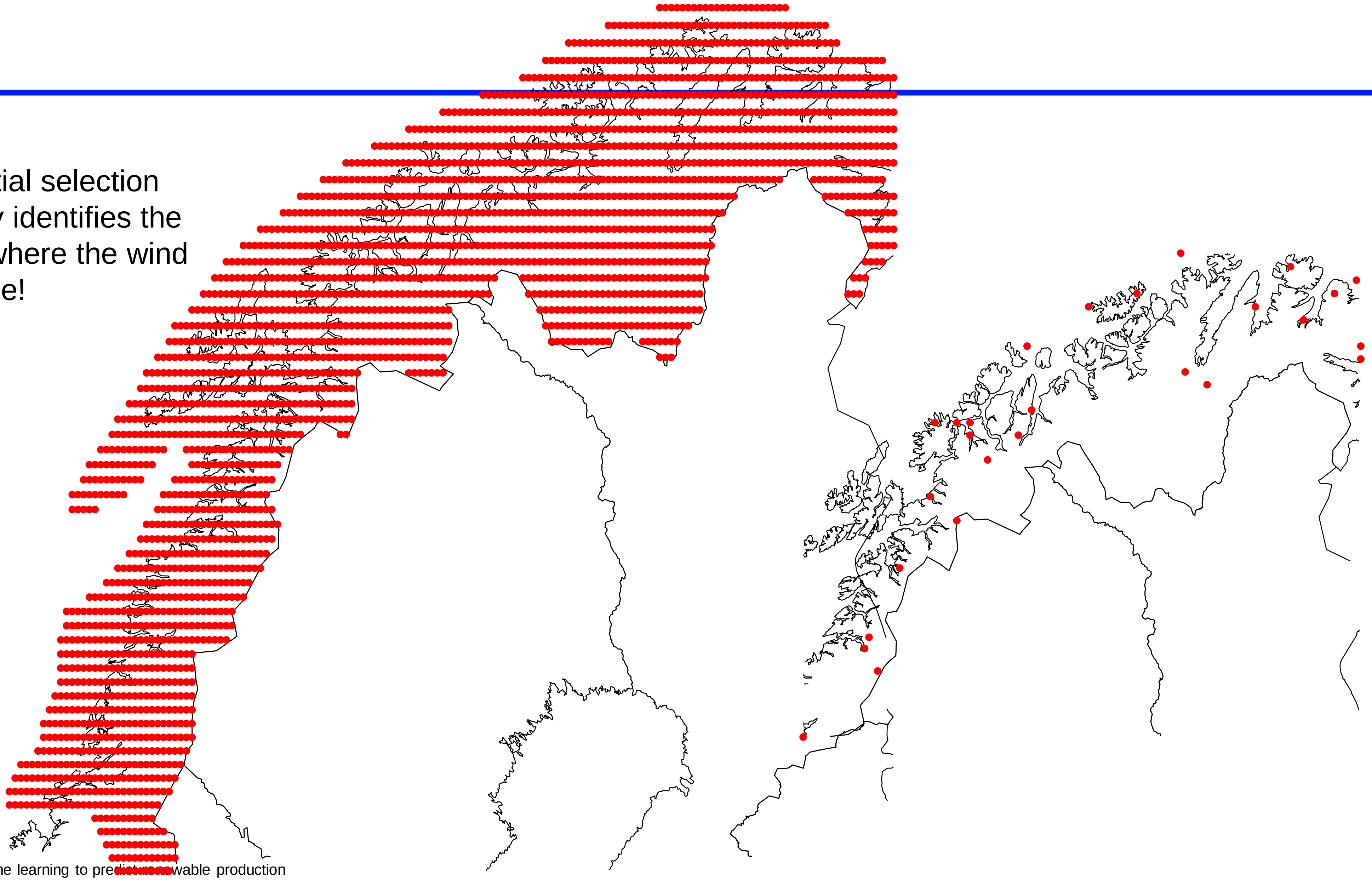


from 1100 to 12 points
(only marginal improvement
in the sequential selection
afterwards…)



Selection by modelling

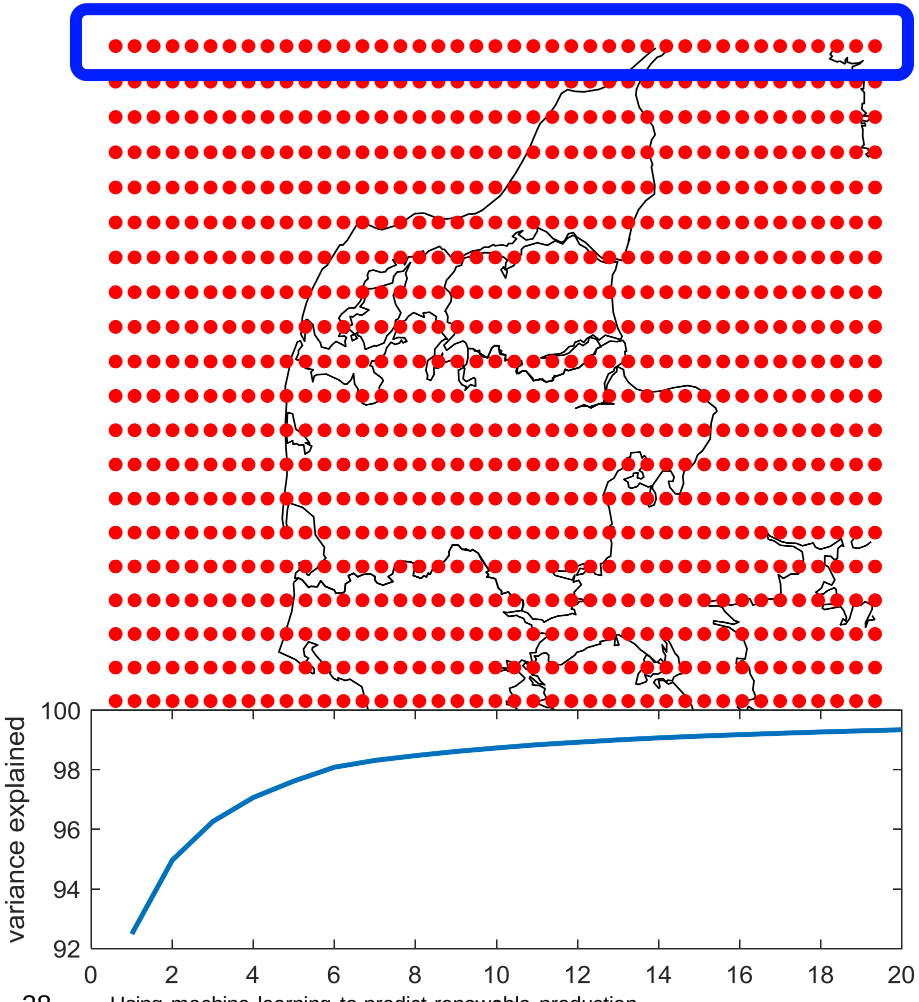
Sequential selection correctly identifies the places where the wind farms are!

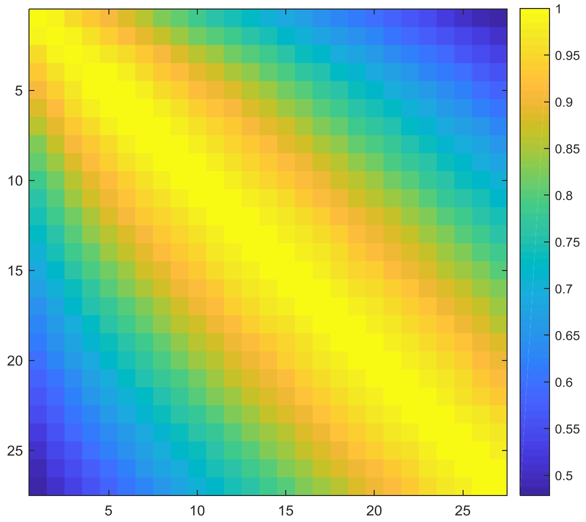


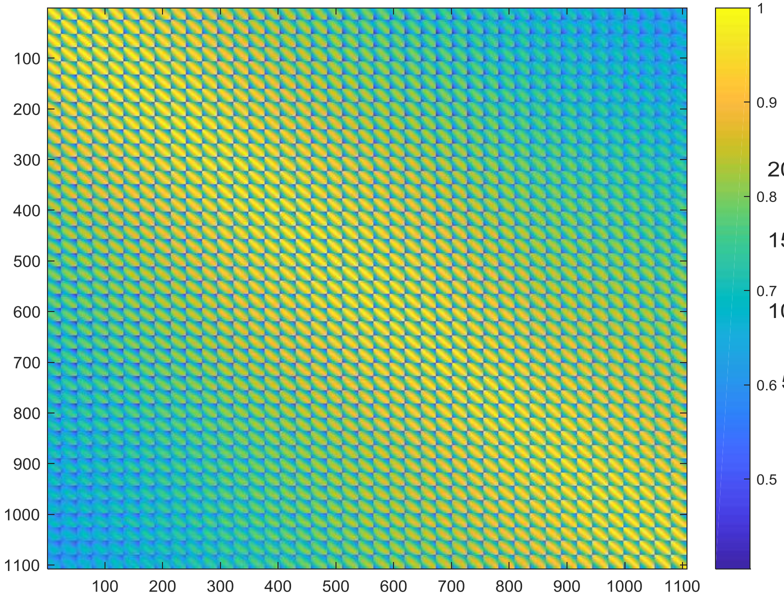
REFINITIV

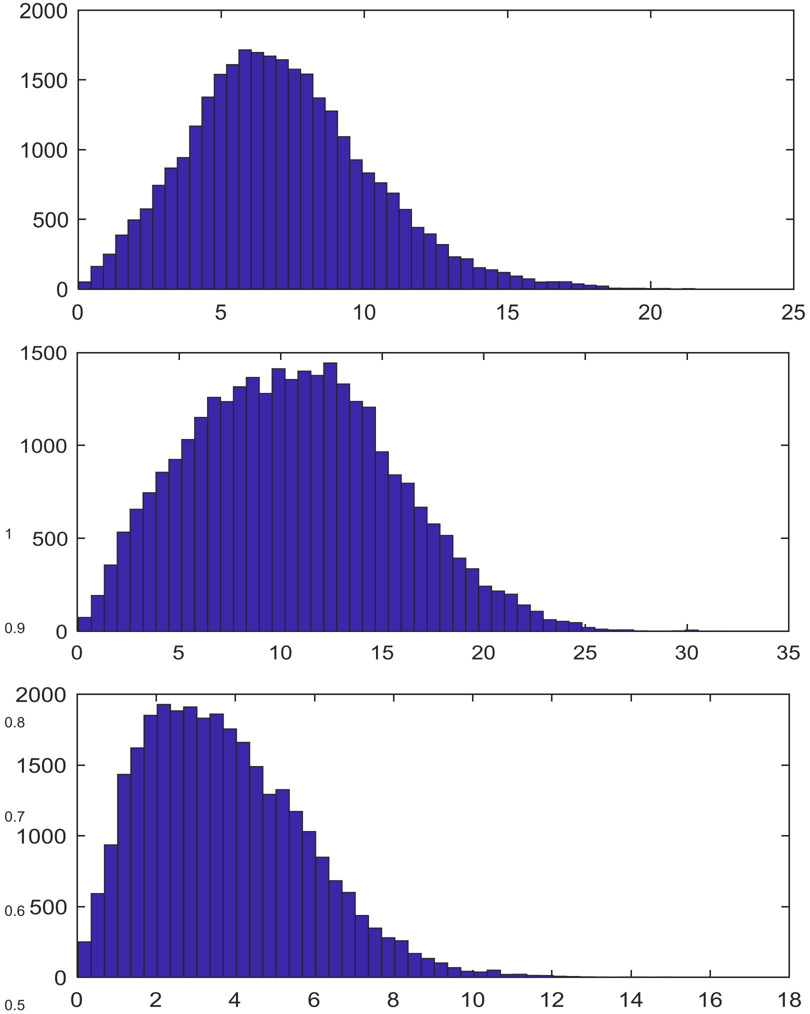
PCA and other variable reduction methods

- Different reasons for PCA not working properly









REFINITIV

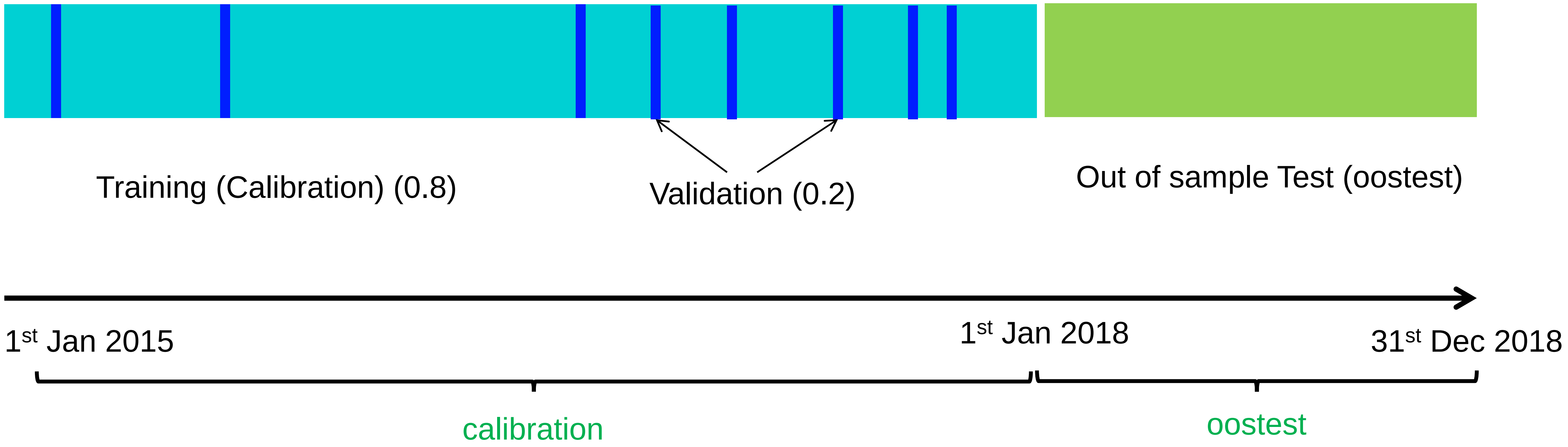
Case studies and results

REFINITIV™

Random forests

Data preparation

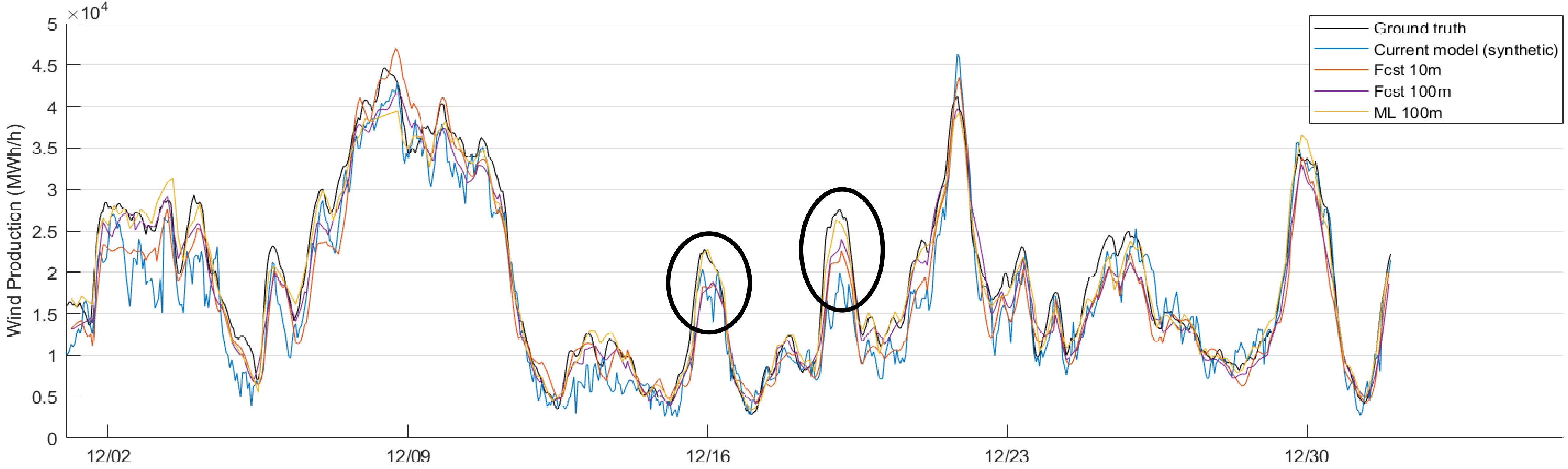
Choice of Training/Validation/Out of sample test datasets



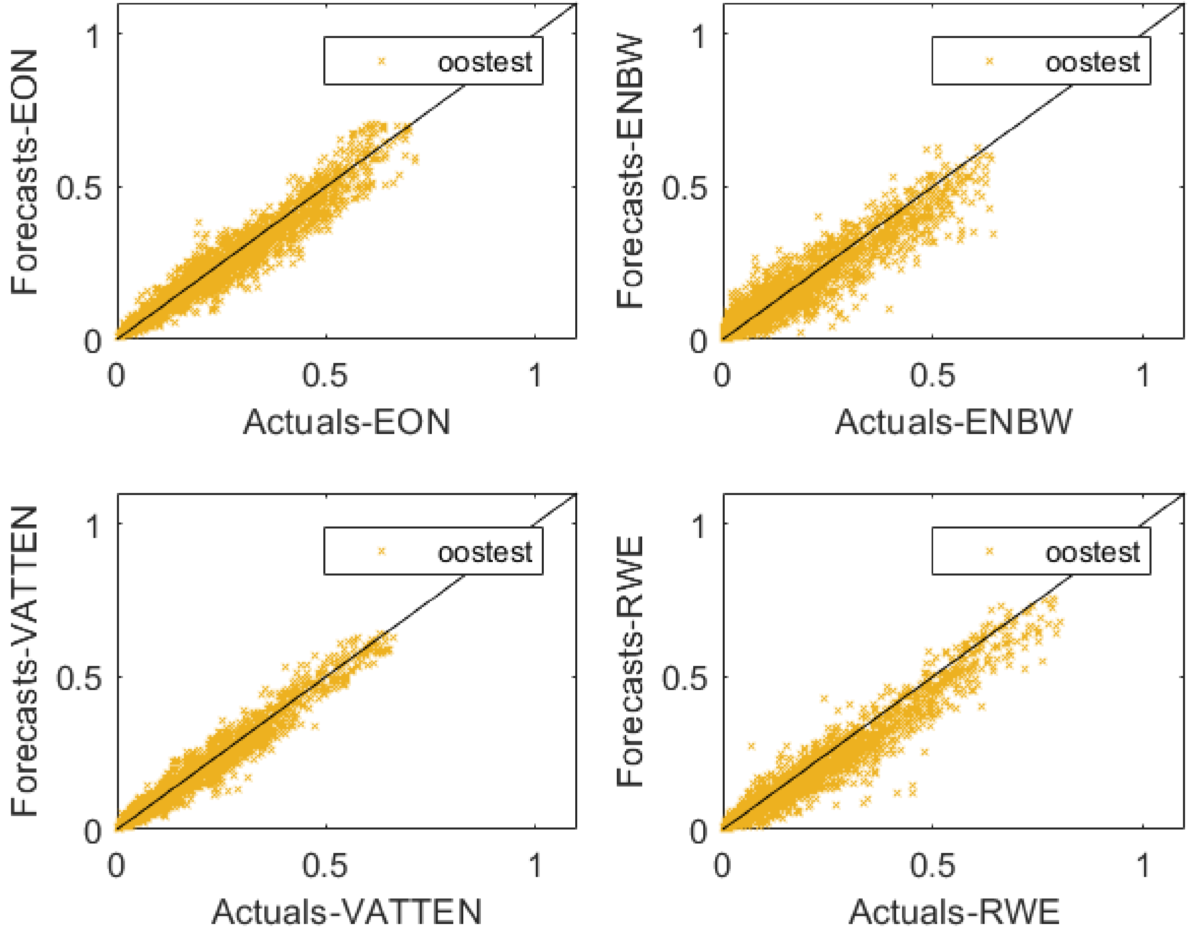
REFINITIV

Results DEU wind

%	Synthetic	Current Day forecast (EC00)	Day ahead forecast (EC00)
Physical model, actual calibration, Fourier expansion	15.0	15.7	17.0
Forecast calibration, 10m height wind	13.0	x	15.6
Forecast calibration, 100m height wind	9.6	x	13.7
ML (Random forest method), 100m height wind	**6.2**	**6.4**	**10.7**
TSO fcst (first coming at 17:00, cont. updating)	3.05 (*)	x	8.1



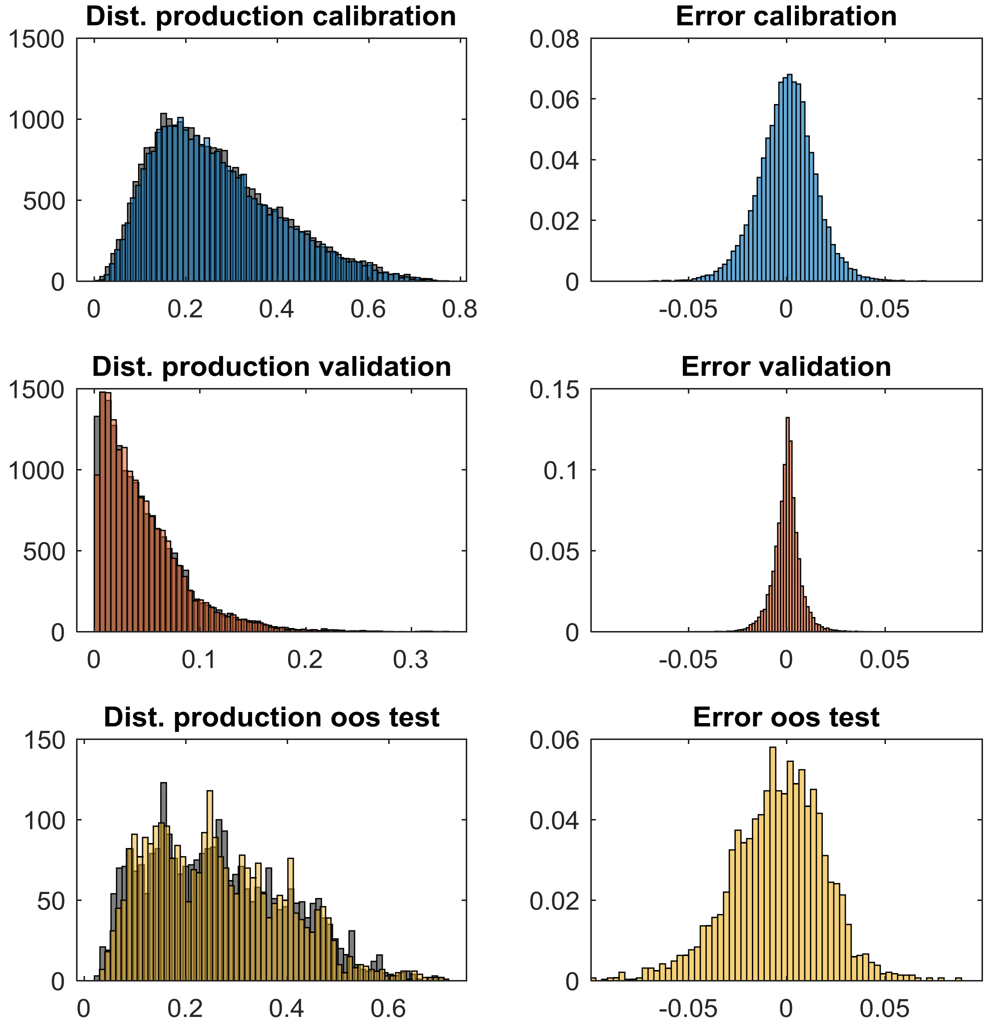
Results DEU wind

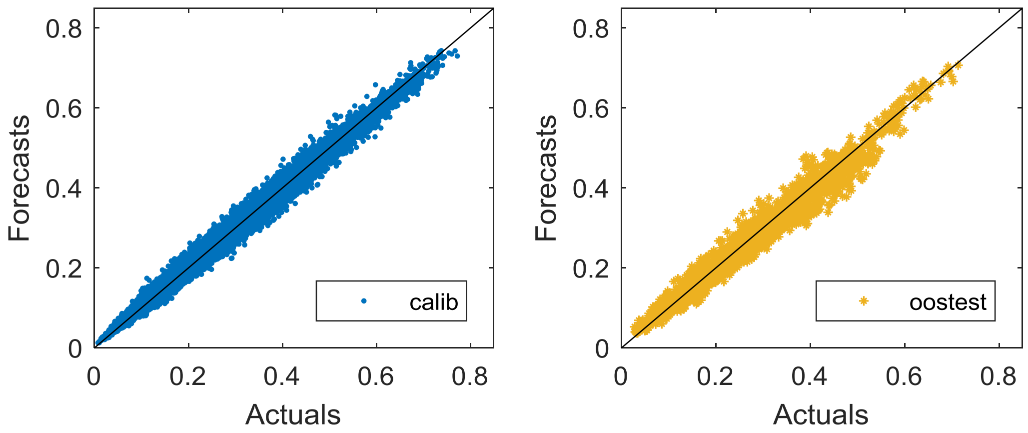


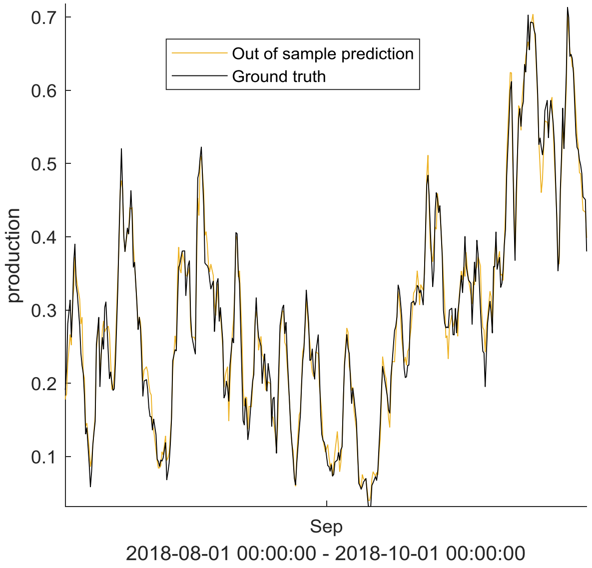
%	Calib	Valid	OOS
EON	5.7	7.5	7.8
ENBW	11.2	12.7	13.2
Vattenfall	6.6	7.5	9.6
RWE	7.4	8.2	8.9
DEU	**4.9**	-	**6.2**

REFINITIV

Results NRD – best setting





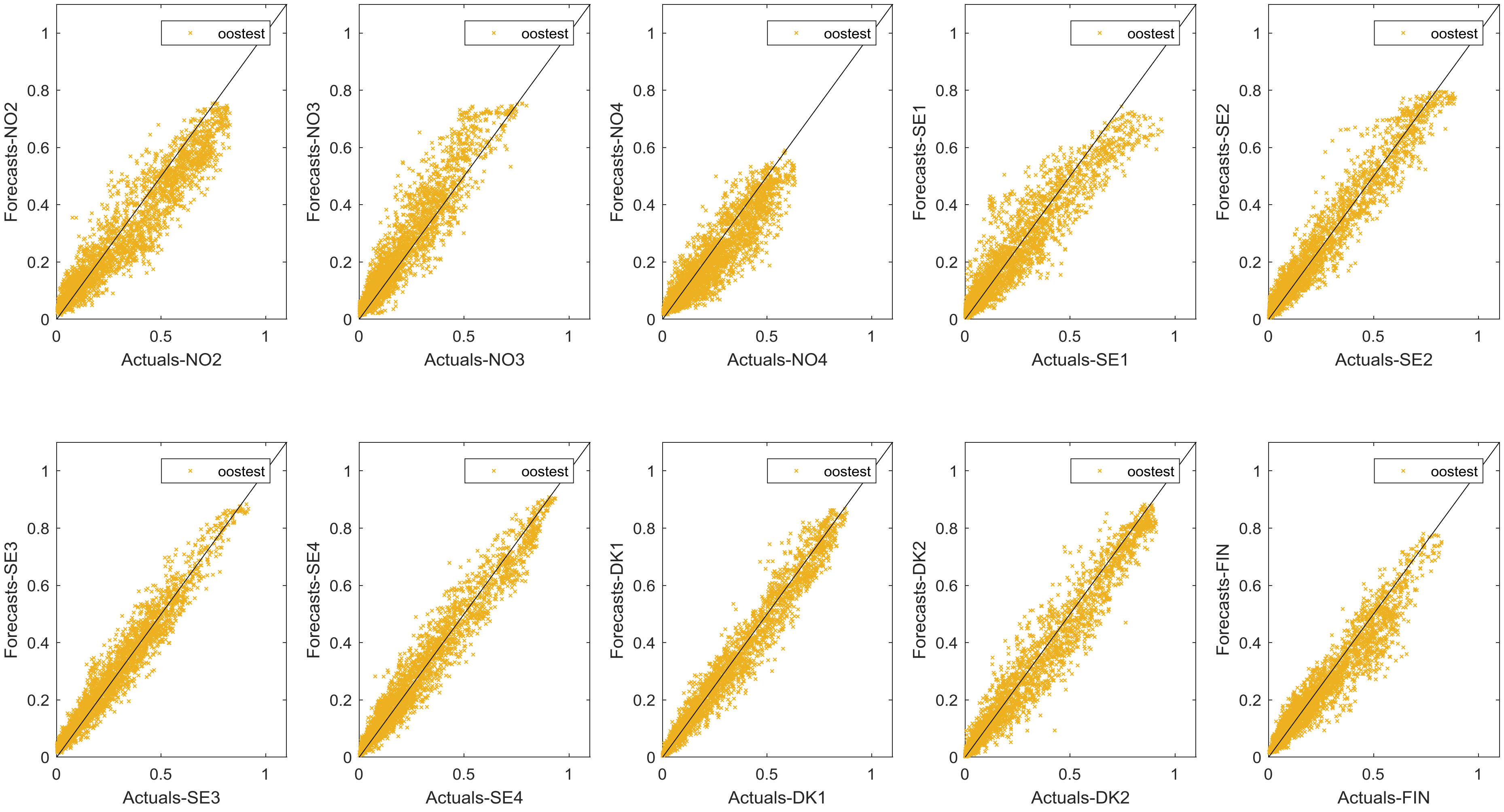


Error:
- Calib 4.0 %
- OOS 6.9 %

Hyperparameters:
- # points per leaf ~15 (!)
- # features per decision split ~20
- max # splits ~ 1000

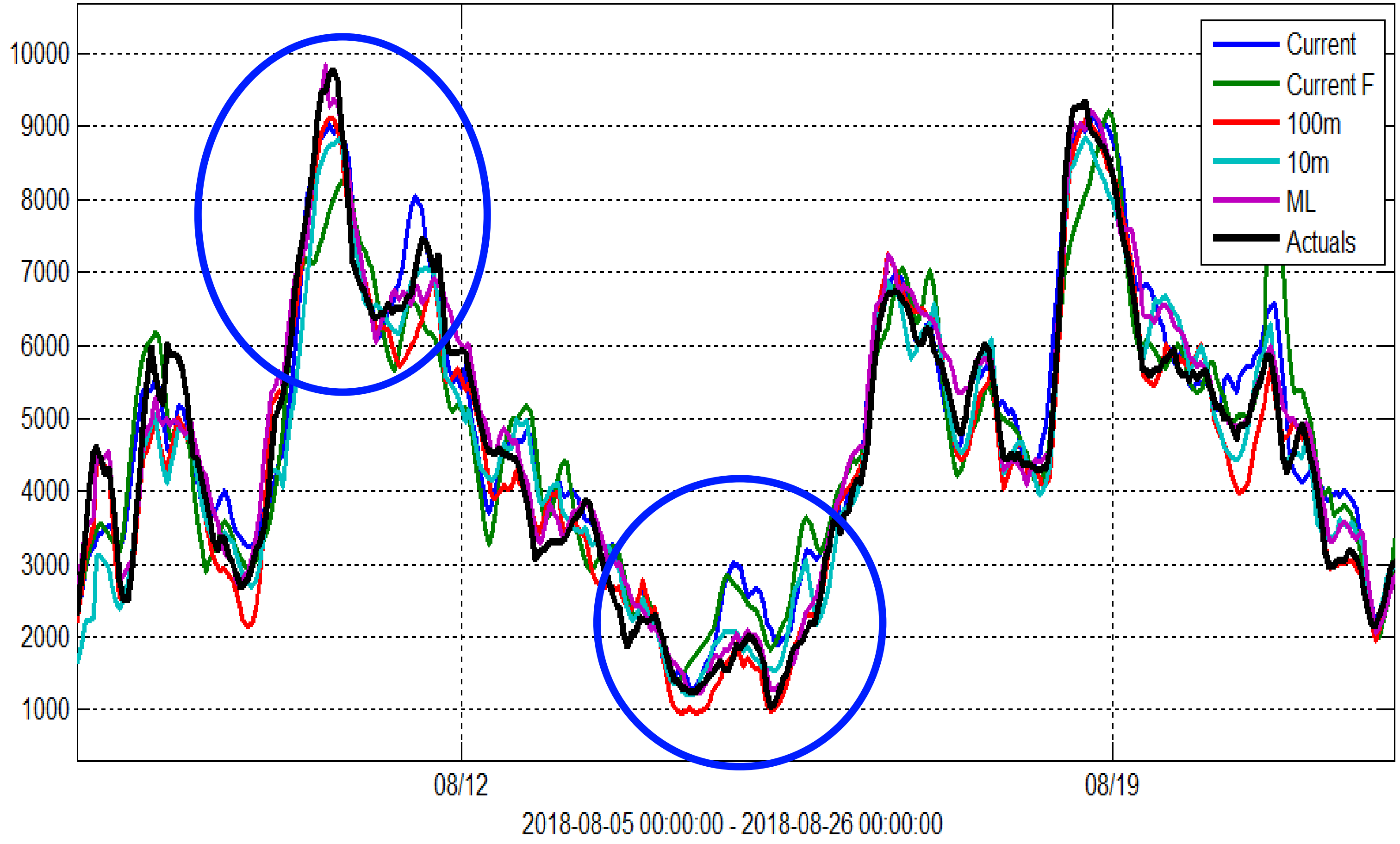
REFINITIV

Results NRD – best setting



%	Calib	Valid	OOS
NO2	13.2	17.3	19.7
NO3	13.2	19.0	25.9
NO4	11.6	15.9	21.9
SE1	10.7	16.4	23.4
SE2	8.9	11.9	16.5
SE3	6.7	9.4	13.3
SE4	8.7	11.6	13.5
DK1	7.9	9.2	11.1
DK2	10.1	12.9	13.6
FIN	8.5	11.4	16.4
Nrd	**4.0**	**-**	**6.9**

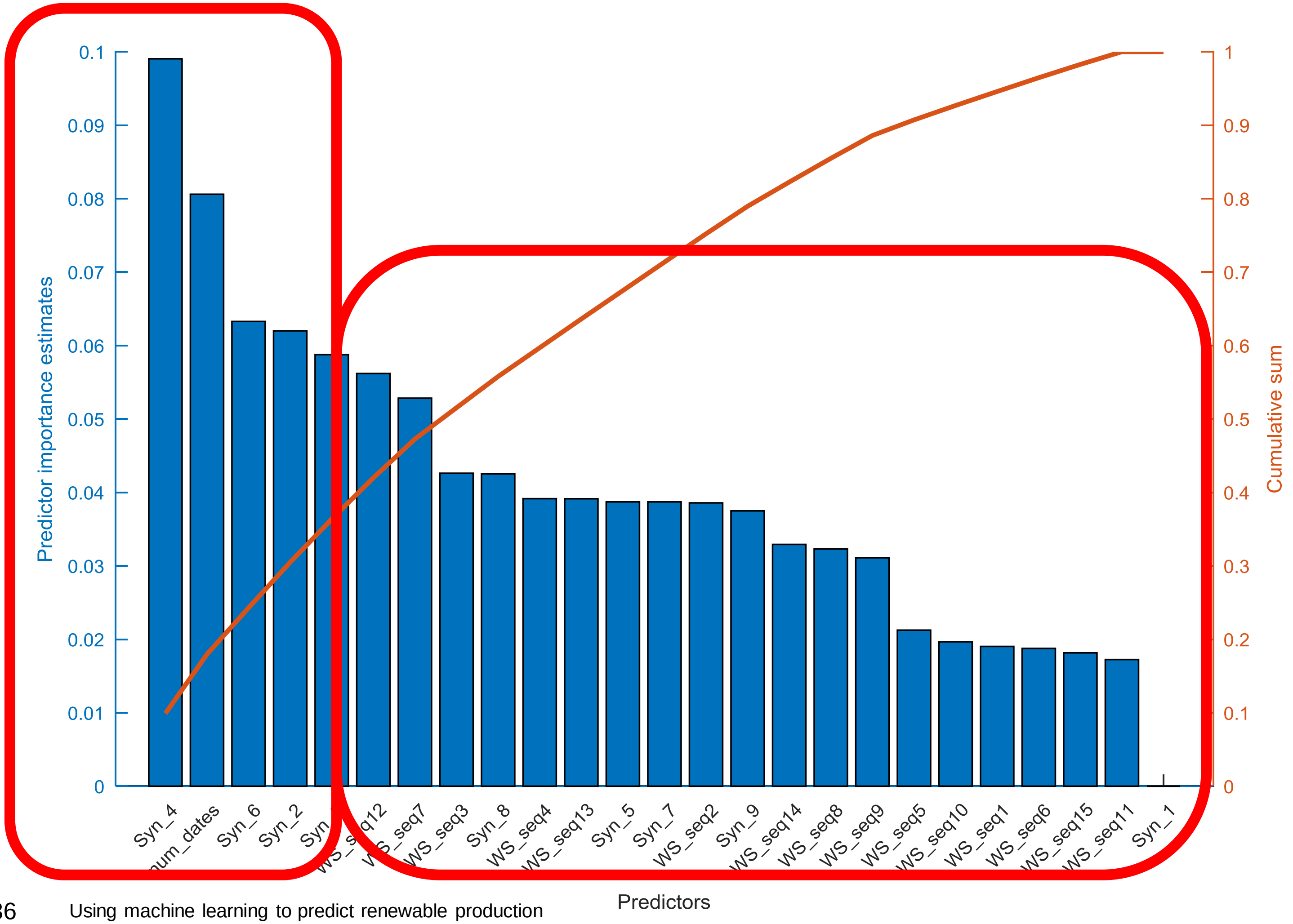
Results NRD, comparison synthetic series on a recent period

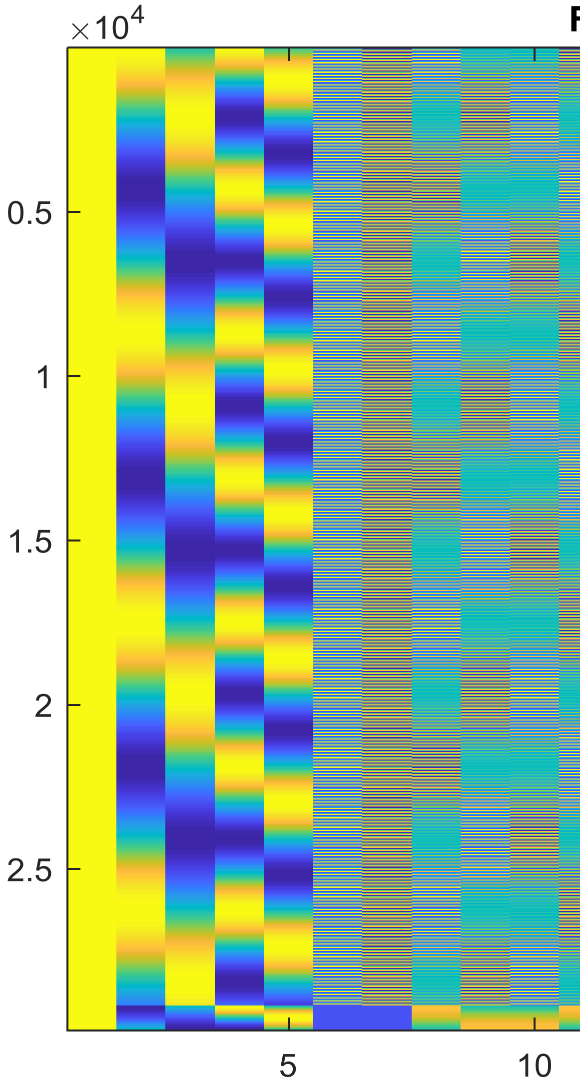


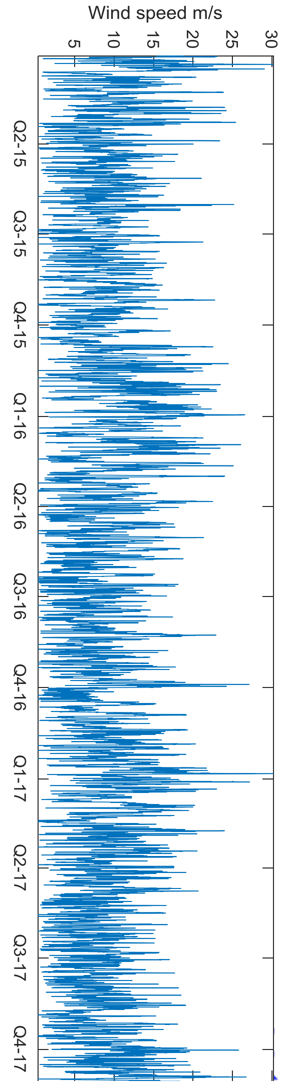
%	Synth	Current Day forecast (EC00)	Day ahead forecast (EC00)
Physical model, calibration on actual wind speed data	12.7	14.9	15.9
Physical model, Fourier expansion	15.5	18.3	19.3
Forecast calibration, 10m height wind	10.3	10.7	12.2
Forecast calibration, 100m height wind	8.0	7.9	10.2
ML (Random forest method), 100 m wind	**7.5**	**7.4**	**9.1**

REFINITIV

Results NRD feature importance

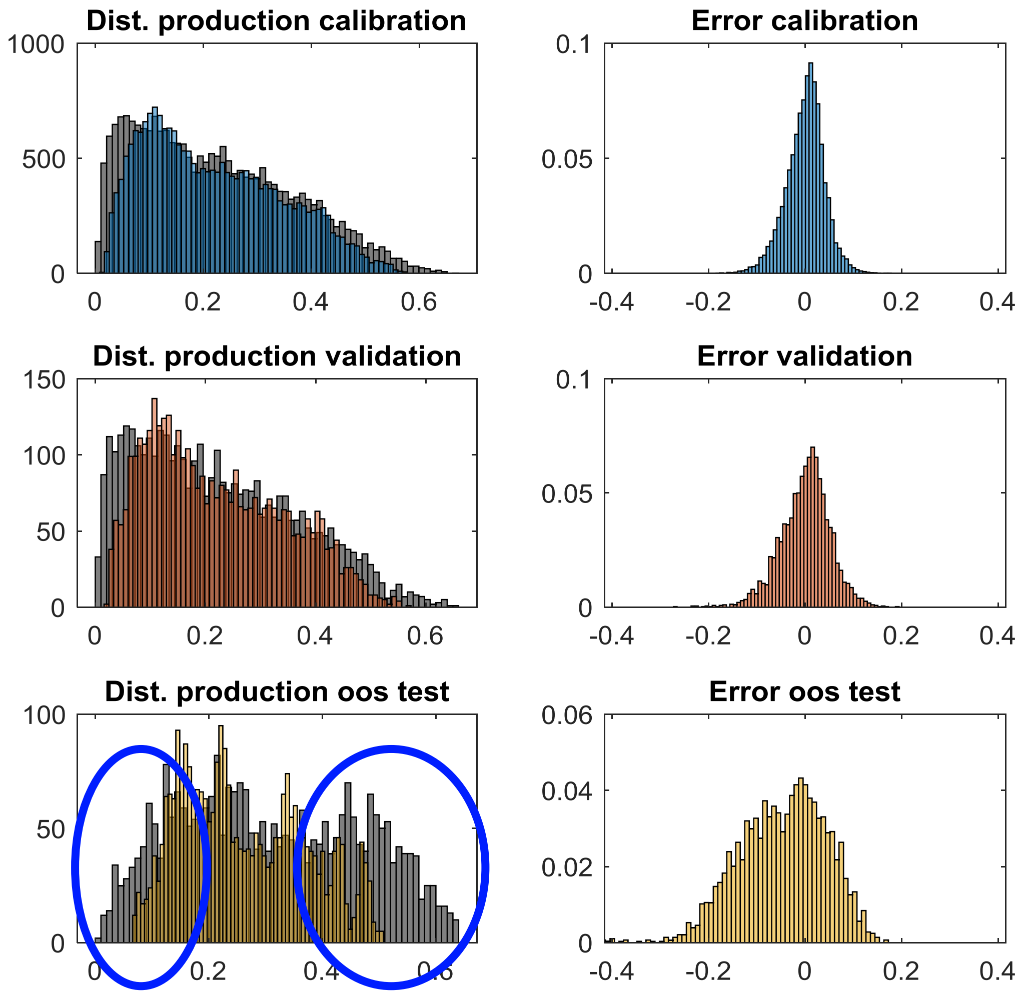




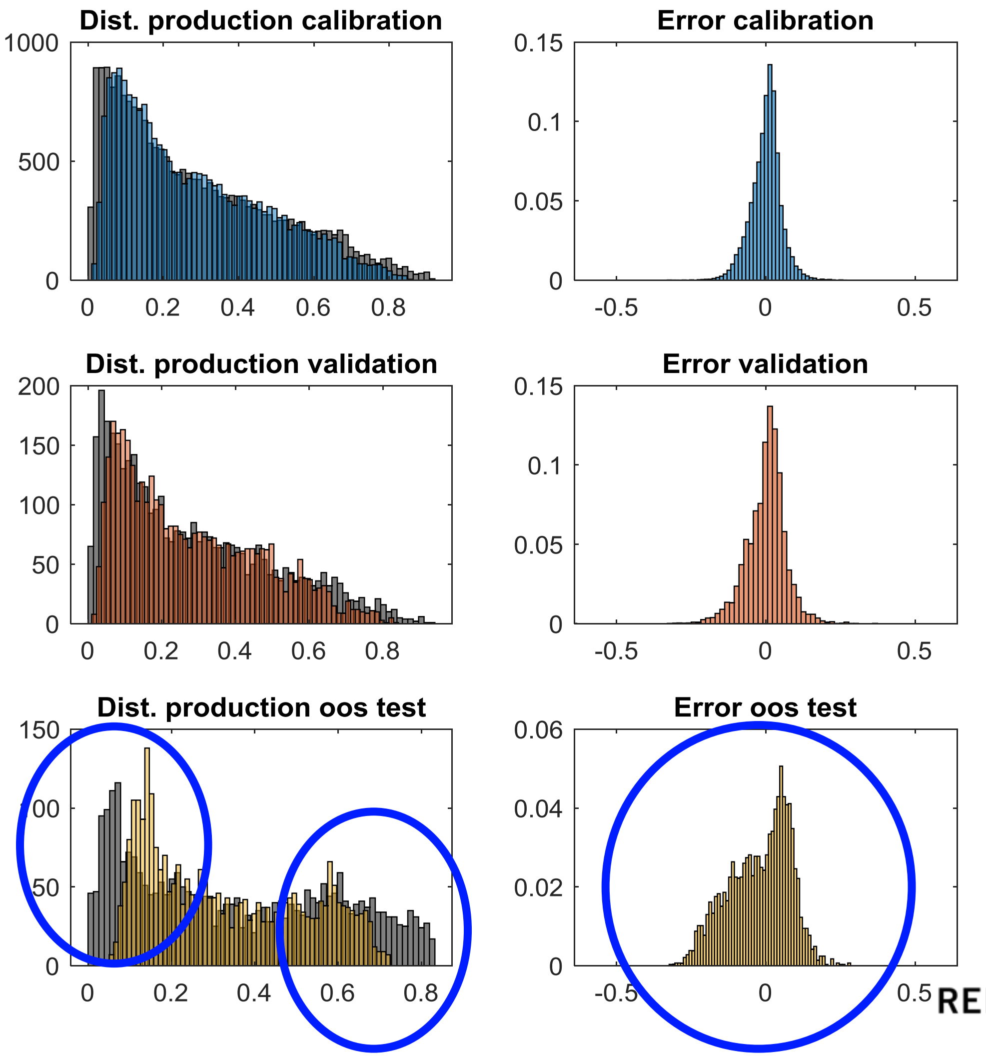


Results NRD, feature selection using PCA

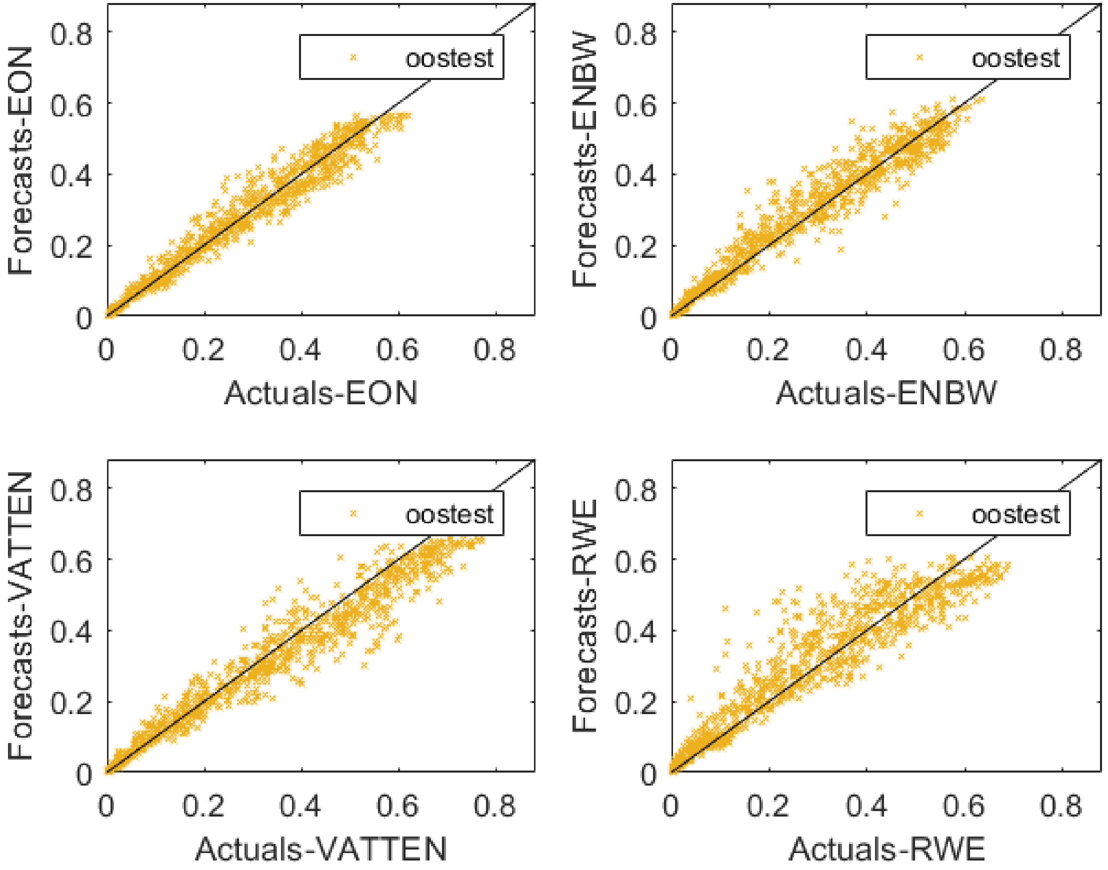
DK1



NO3

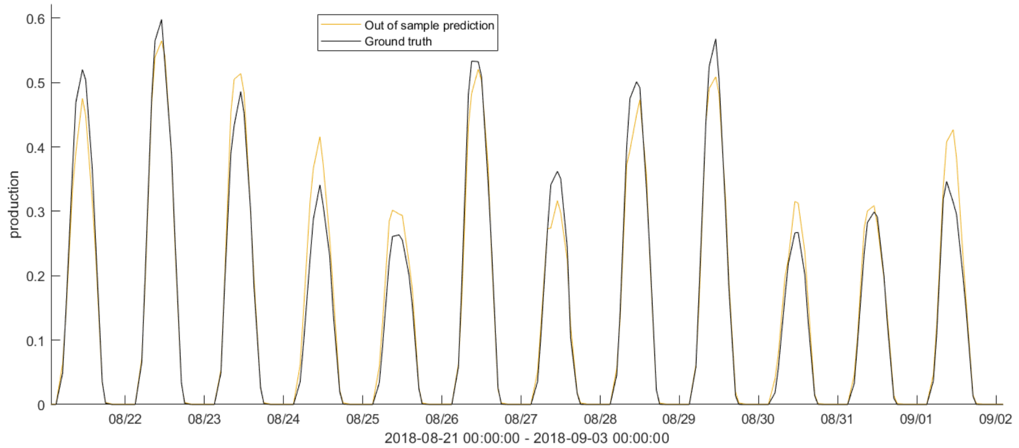


Results DEU solar (PV)



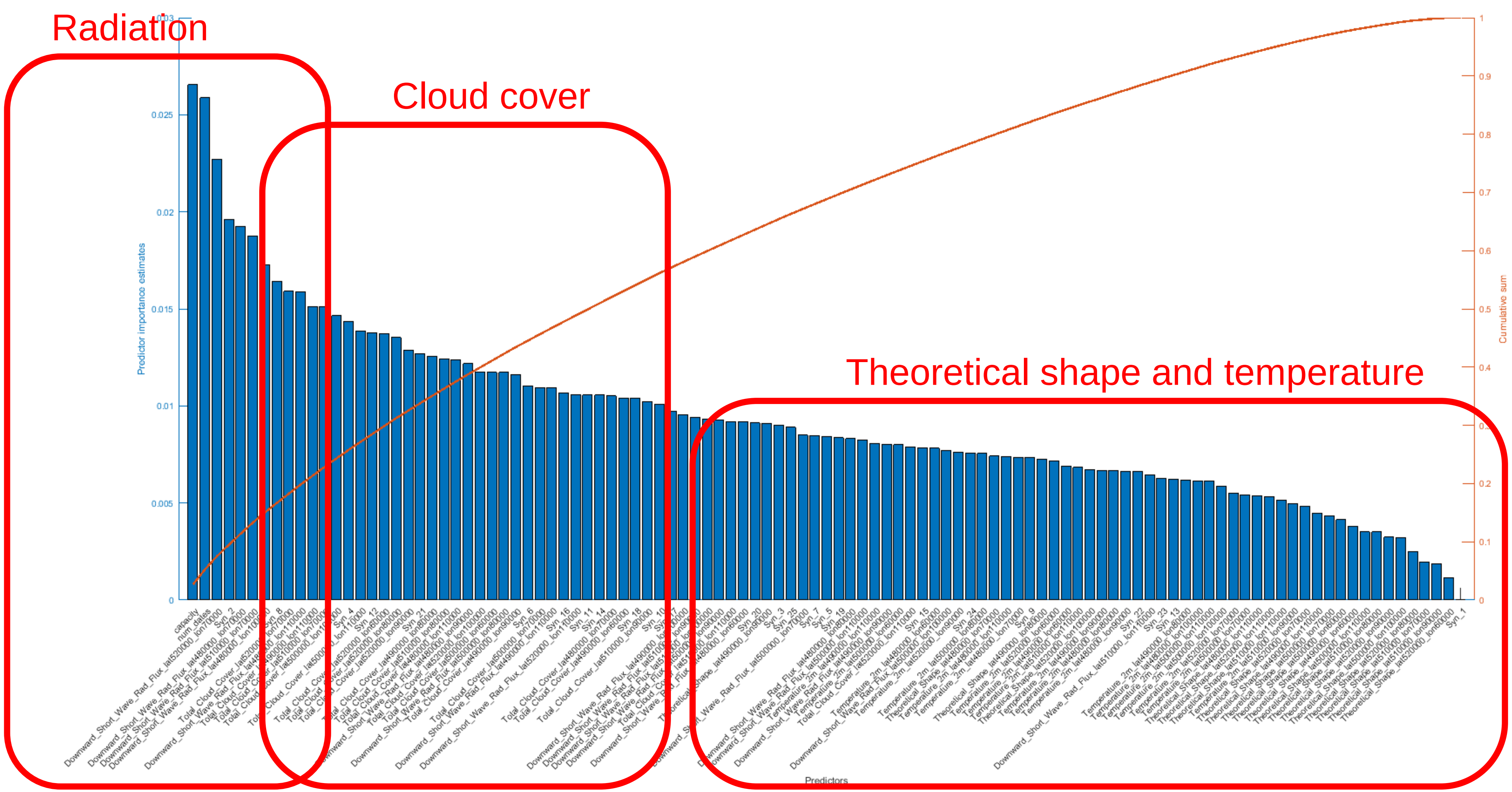
Aggregated

%	Calib	Valid	OOS	Curr synth
EON	9.9	-	**10.2**	11.9
ENBW	11.0	-	**11.2**	13.4
Vattenfall	10.3	-	**12.1**	15.4
RWE	12.1	-	**18.1**	17.2
DEU	**8.3**	**-**	**8.6**	**8.2**



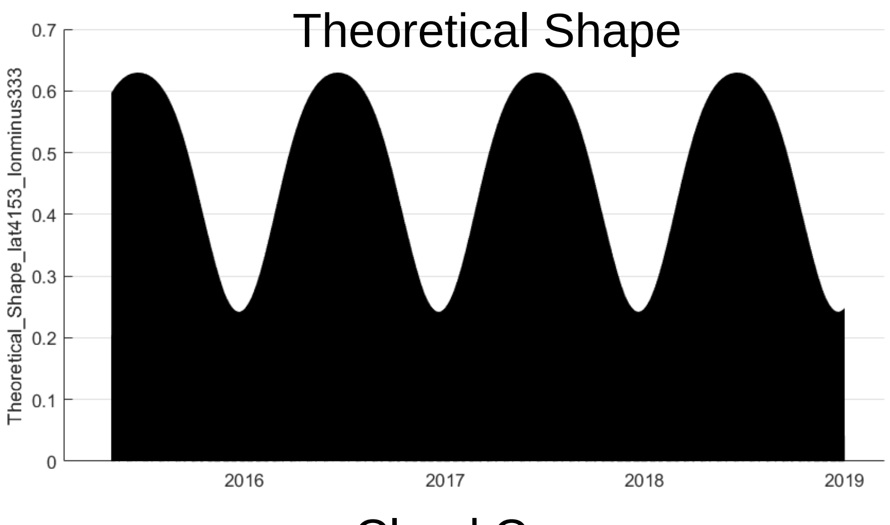
REFINITIV

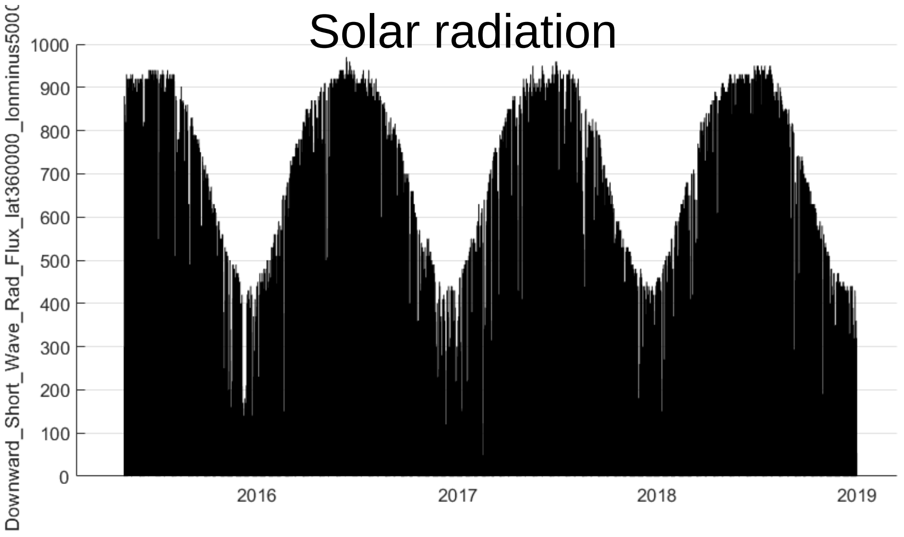
Results DEU solar – feature importance

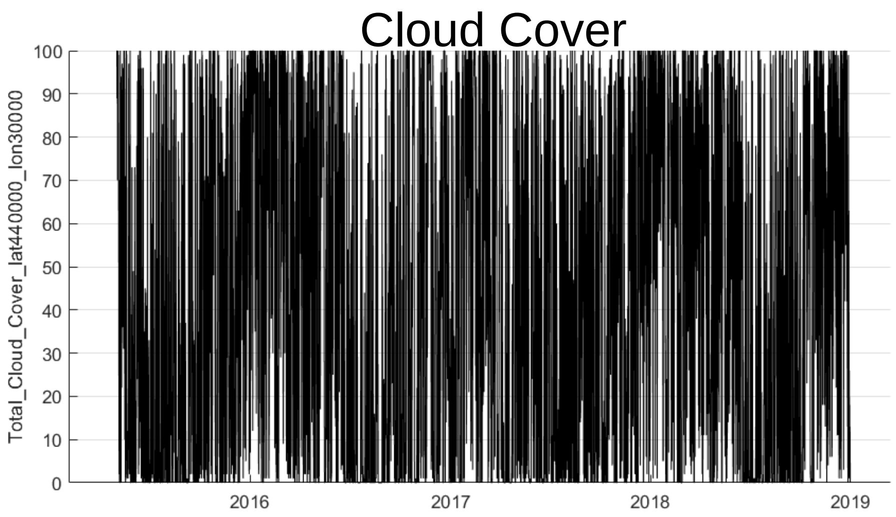


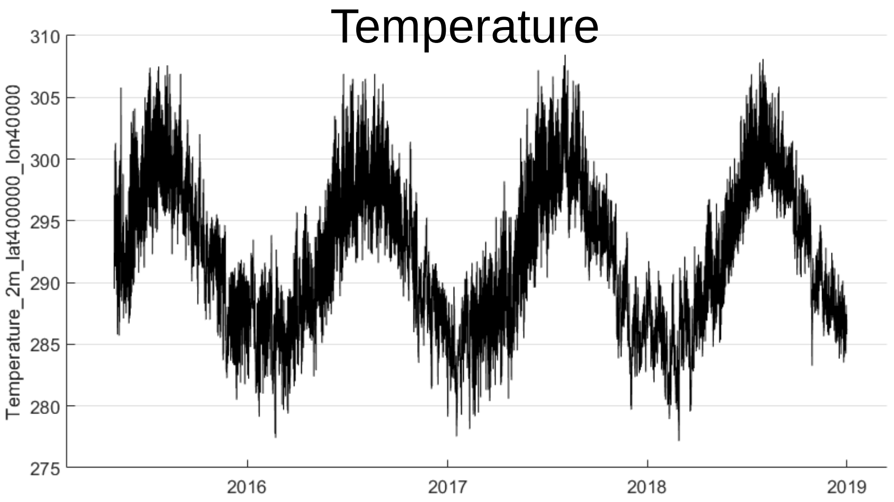
REFINITIV

Results ESP PV solar: input features





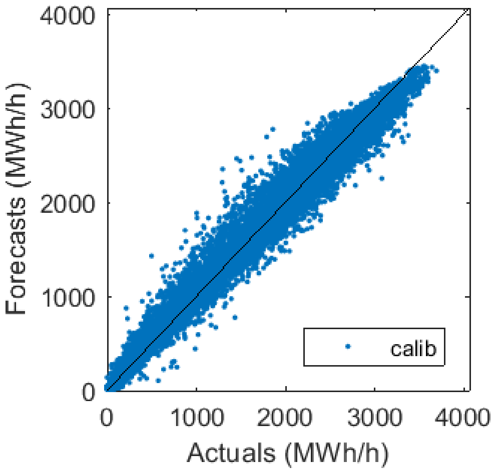


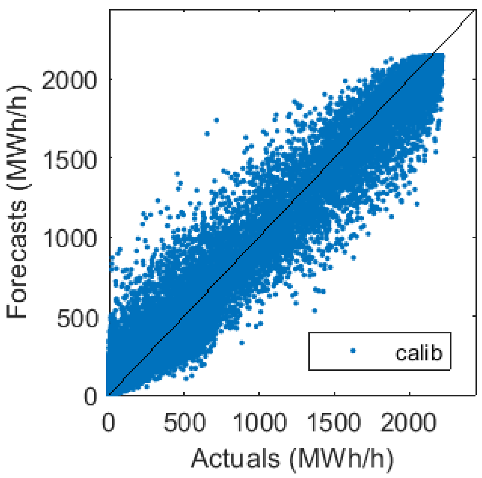


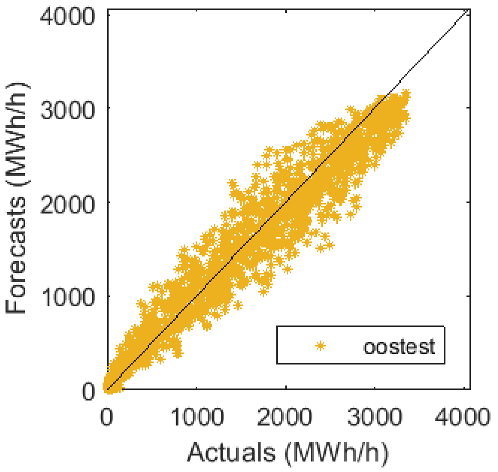
Results ESP solar

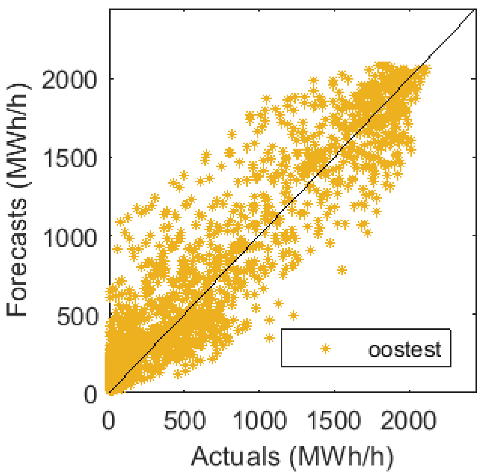
PV

Thermal



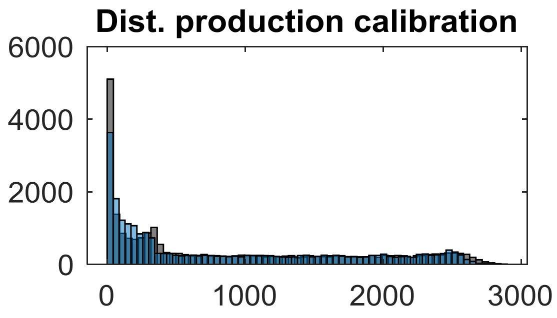


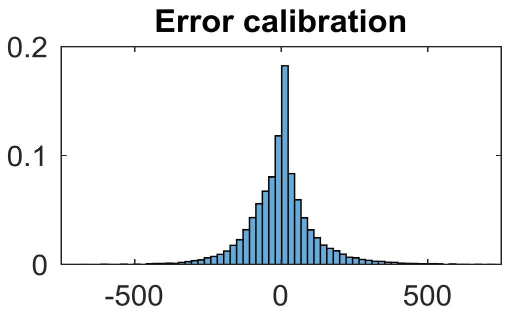


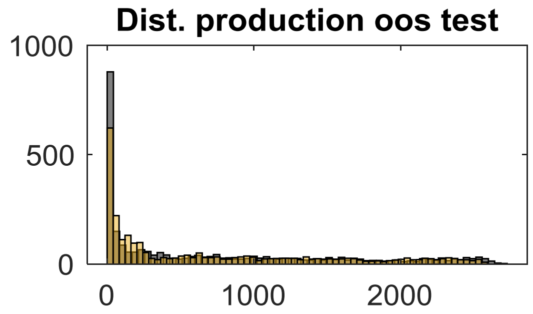


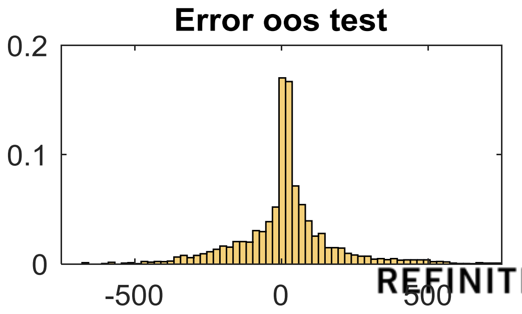
Aggregated

%	Calib	Valid	OOS	Curr synth
PV	6.7	8.0	**12.4**	**13.3**
Thermal	15.2	18.9	**28.3**	x
ESP	**8.9**	**12.6**	**15.2**	



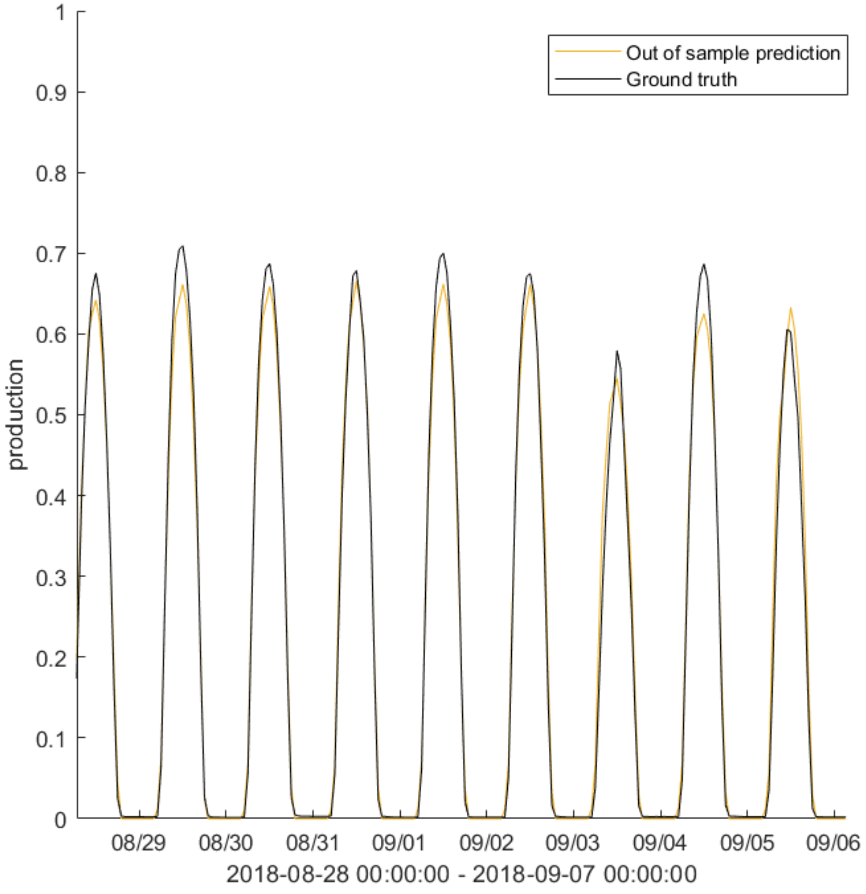




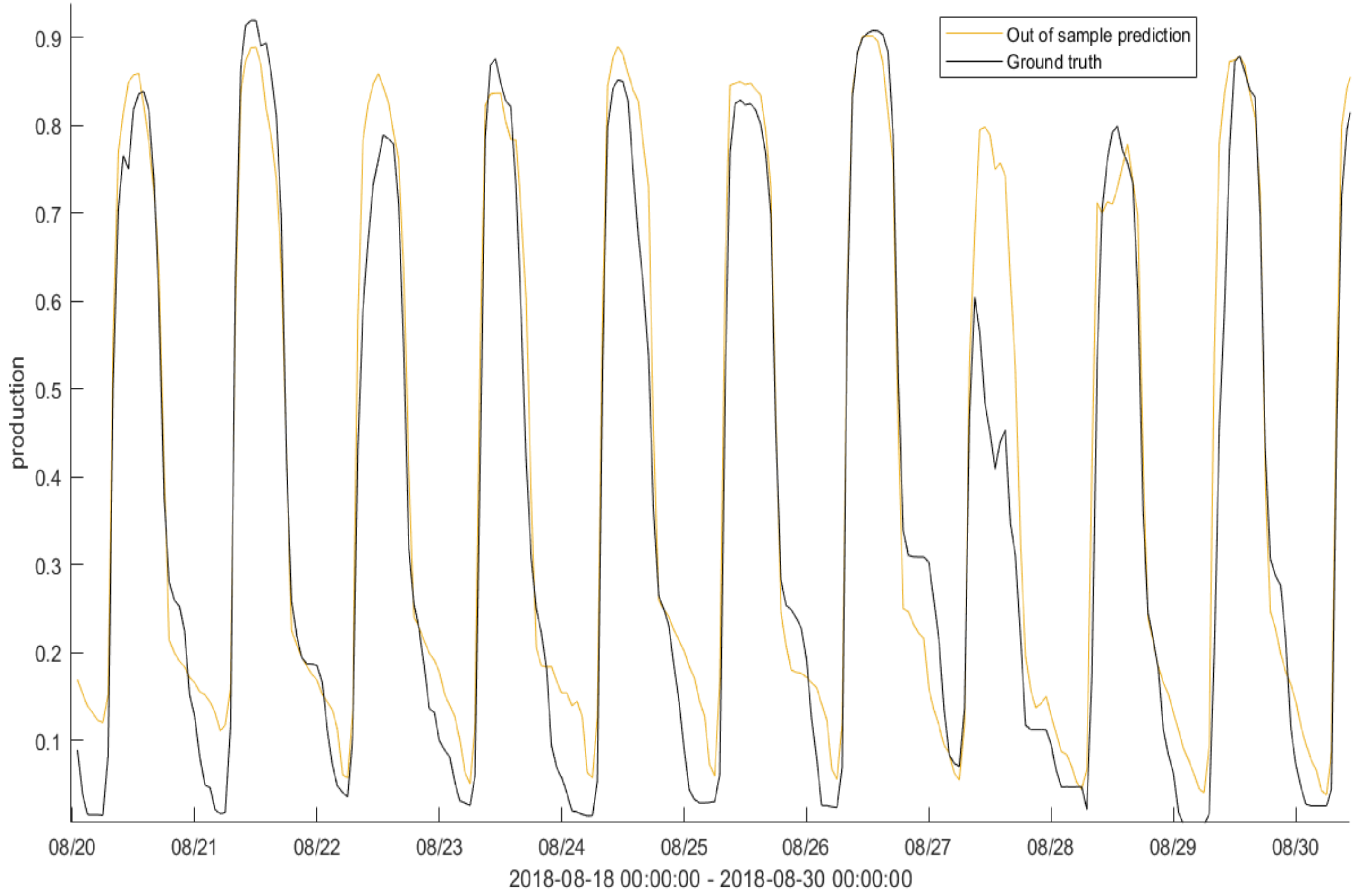


Results ESP solar

PV



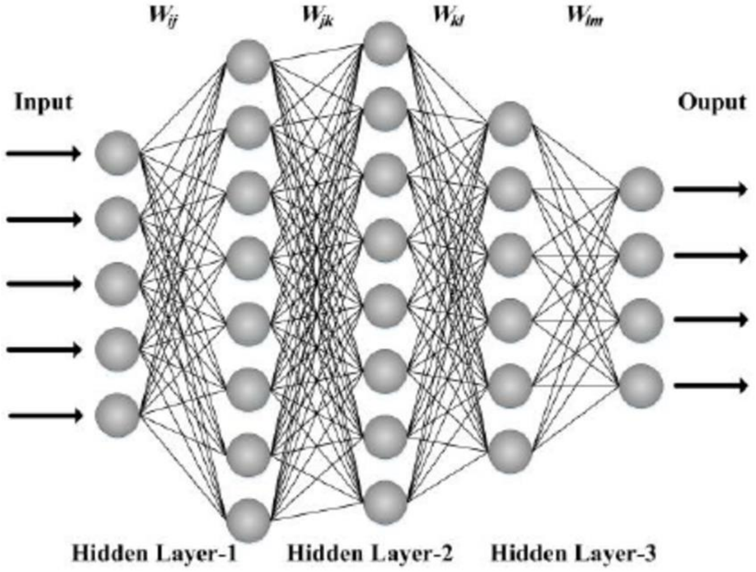
Thermal



Other ML methods tested

Neural Networks

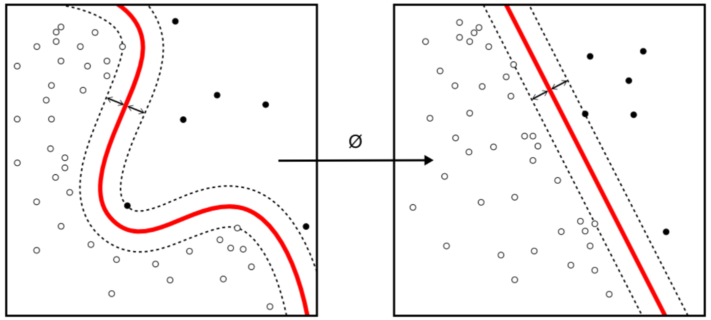
A layered network of basic functions able to represent the complex relationship between input and output



Similar results on fundamental prediction, larger out-of-sample error wrt calibration error (overfitting)

SVM

A support-vector machine constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, regression or other tasks like outliers detection.



In general, in our experience, Random Forests tend to perform better than SVM for **regression** problems.

Naïve explanation Random Forest is intrinsically suited for multiclass problems, while SVM is intrinsically two-class

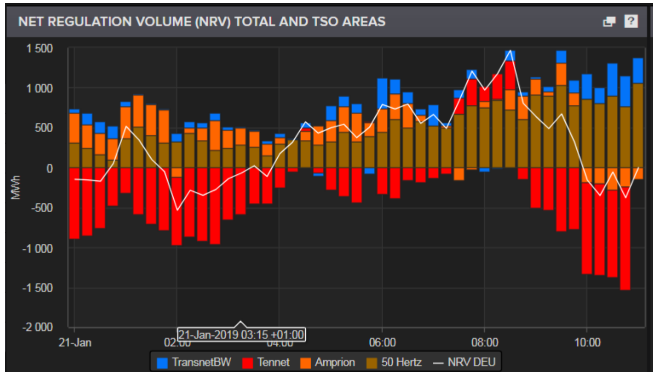
Truth Random Forest works well with a mixture of numerical and categorical features. When features are on the various scales, it is also fine.

REFINITIV

Conclusions

Another project where we are testing ML techniques:

- Intraday balancing volume forecasts using LSTM (long short-term memory) NN



Output:
Imbalance volume (15min)

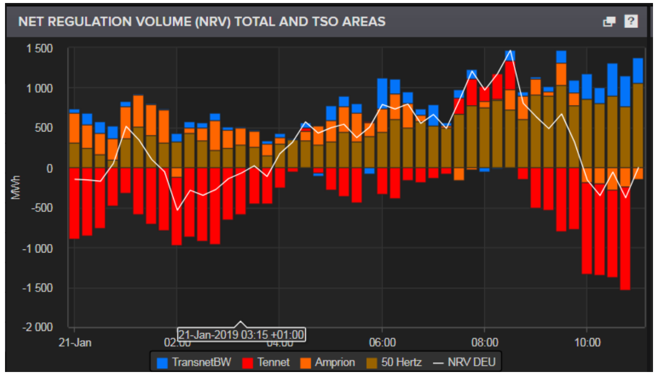
- Δ wind forecast to wind actual
- Δ solar forecast to solar actual
- Δ consumption forecast to actual
- Changes in UMMs / availabilities
- Outages
- Change in unregulated inflow
- ....
- ....

features

Conclusions

Another project where we are testing ML techniques:

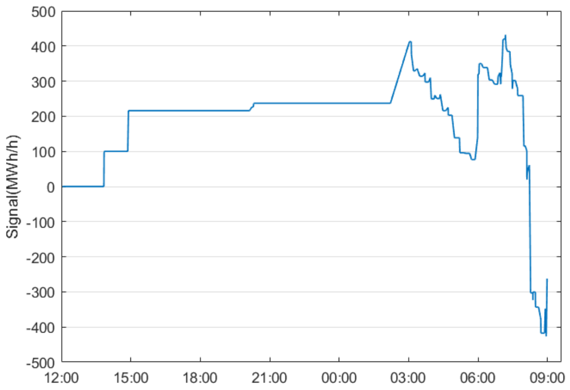
- Intraday balancing volume forecasts using LSTM (long short-term memory) NN



Output:
Imbalance volume (15min)

LSTM / Timeseries approach*:*
Each feature has an evolution through time until delivery

- Δ wind forecast to wind actual
- Δ solar forecast to solar actual
- Δ consumption forecast to actual
- Changes in UMMs / availabilities
- Outages
- Change in unregulated inflow
- ....
- ....



features

REFINITIV

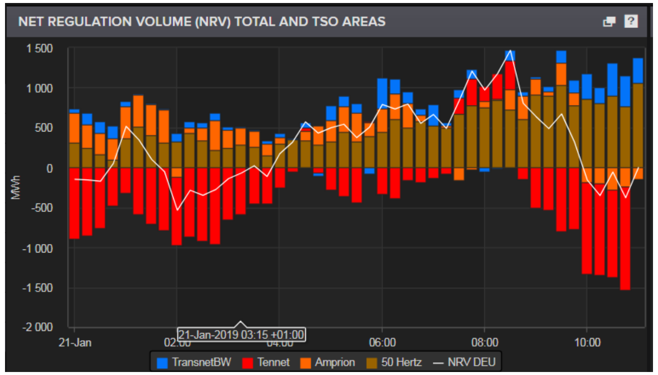
Conclusions

Another project where we are testing ML techniques:

- Intraday balancing volume forecasts using LSTM (long short-term memory) NN

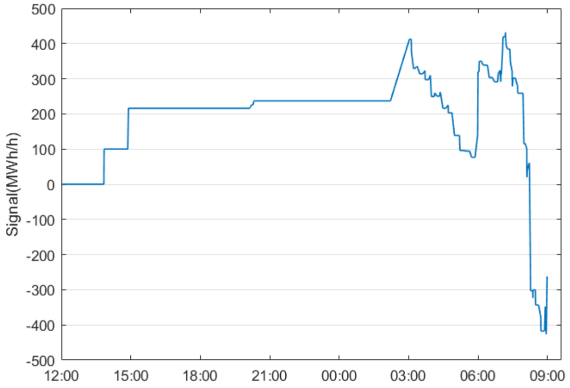
Lessons learned:

- Ability of handling big weather data in ML models depends on having good routines for:
    - Data processing and storing
    - Variable selection (dimensionality reduction)
- Sensible improvemements in OOS (out-of-sample) error thanks to calibration on forecast grid and smart data handling
- Fine tuning of hyperparameters to avoid overfitting
- Harder to beat current stack based or bid-offer based models in pure short term price modelling



Output:
Imbalance volume (15min)

LSTM / Timeseries approach:
Each feature has an evolution through time until delivery



- ∆ wind forecast to wind actual
- ∆ solar forecast to solar actual
- ∆ consumption forecast to actual
- Changes in UMMs / availabilities
- Outages
- Change in unregulated inflow
- ....
- ....

features

Thank you

gabriele.martinelli@refinitiv.com

REFINITIV™