Report 03/05

May 6, 2024

Introduction

This presentation shows a summary of the work done in the past few weeks. It includes the models used and the results obtained.
This document is for internal use, so it may contain some errors.

Models

- Naive (nv)
- Naive 2.0 (nv2)
- KNNR + GA algorithm (knnr)
- KNN regression (knnreg)

Naives

Naive:

$$y_{ih} = \beta_0 + \beta_{1h} + \beta_{2m} + \beta_3 * avg\_sfcWind + \beta_{4h} * avg\_sfcWind + \epsilon_i \quad (1)$$

Naive 2.0:

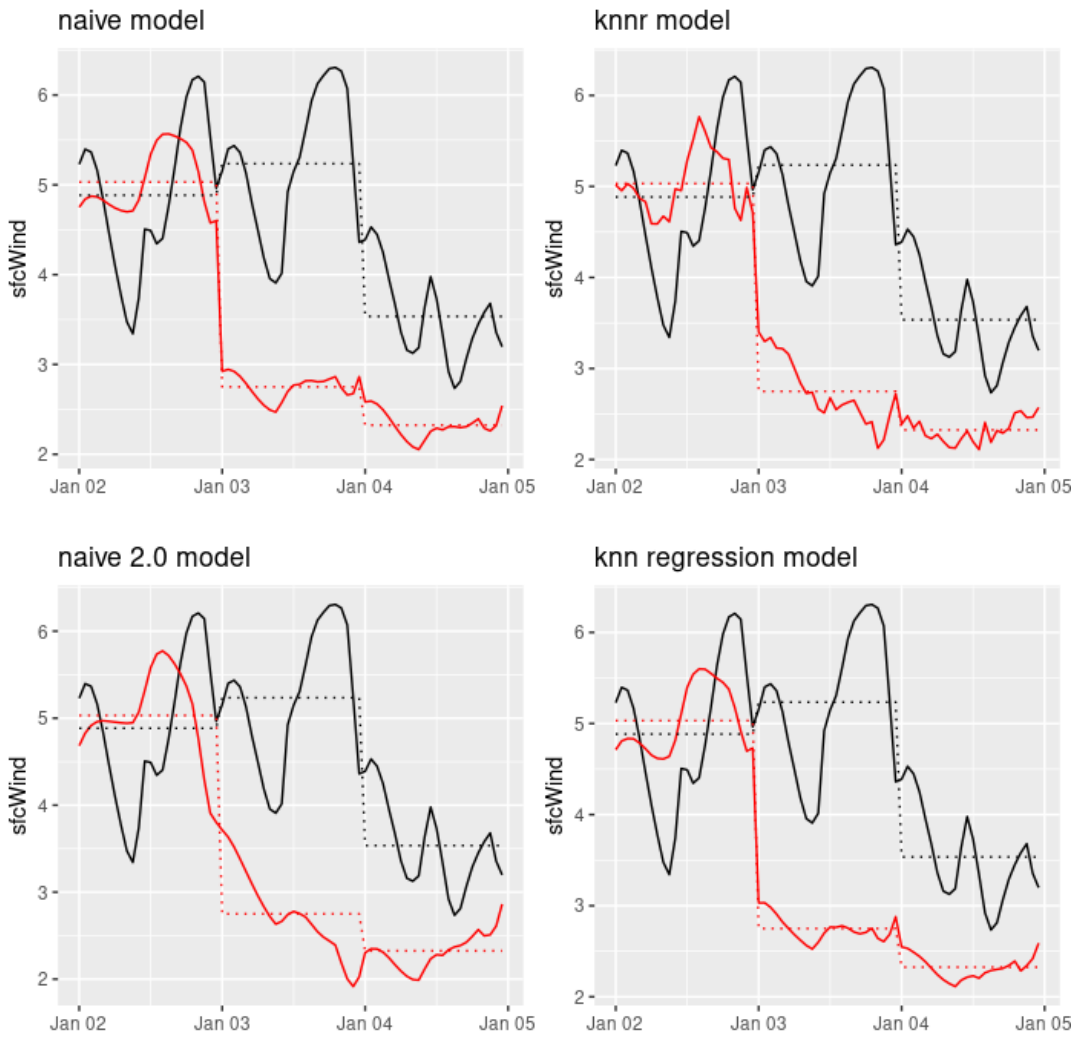
Adds to Equation 1 the following terms:

$$\begin{aligned} &\beta_5 * prev\_avg\_sfcWind + \beta_6 * nxt\_avg\_sfcWind \\ &+ \beta_{7h} * prev\_avg\_sfcWind + \beta_{8h} * nxt\_avg\_sfcWind + \epsilon_i \end{aligned} \quad (2)$$

Other algorithms

- KNNR + GA algorithm (knnr)
  - We implement the algorithm showed in Taesam Lee and Changsam Jeong (2014) paper.
  - As we are using a GA algorithm, it's necessary to run the algorithm many times to get a stable result. We run the algorithm 10 times. The probability of crossover was 0.3. We need to discuss about the mutation step.
- KNN regression
  - We don't adjust the hyperparameters. Number of neighbors and the weight function were fixed. Also we don't use the month as a possible predictor.

Series


naive model
knnr model
naive 2.0 model
knn regression model
sfcWind
2
3
4
5
6
Jan 02
Jan 03
Jan 04
Jan 05

Comments

- There is a difference between the average value of the day of the real and the cmip data.
- We can see that serie has a different behavior in the reanalysis data and the downscaled data.
  - In all the models the amplitude of the series seems smaller than the amplitude of the reanalysis data.
  - The models seem to have a bias in the prediction of the peaks (and valleys).

	diff_of_means	ratio_of_sd	ks_test	amp_rtio_means	max_error	sign_cor
nv	0.240	0.840	0	0.322	0.488	0.265
knnr	0.240	0.861	0	0.554	0.210	0.040
nv_2	0.240	0.863	0	0.523	0.344	0.175
knnreg	0.236	0.828	0	0.342	0.313	0.205

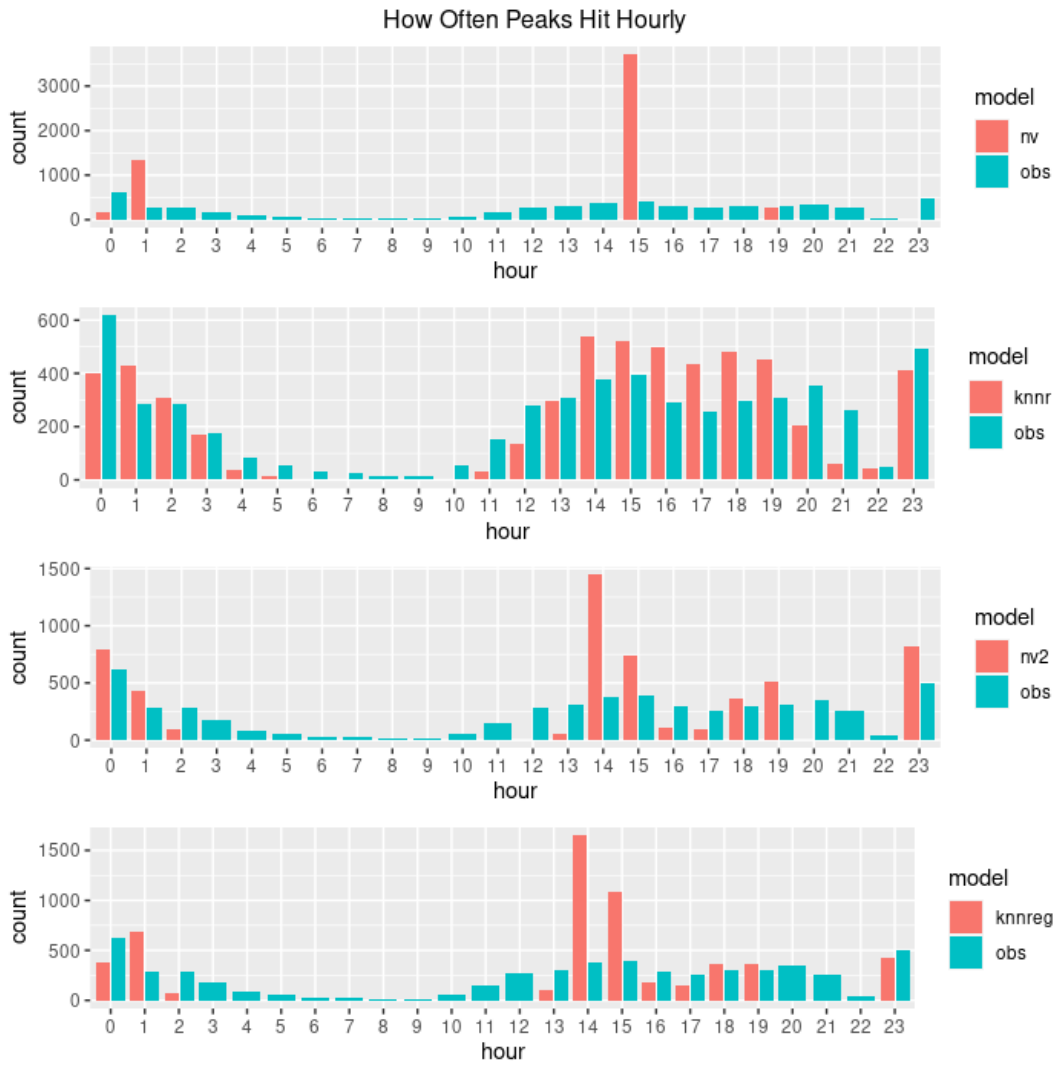
Comments 1/2

- $diff\_of\_means = \bar{Y} - \bar{\hat{Y}}$
- $ratio\_of\_sd = \frac{\sigma_{\hat{Y}}}{\sigma_Y}$
- ks_test is the p-value of the Kolmogorov-Smirnov test
- $amp\_rtio\_means = \frac{\sum_{d=1}^{n} max(\hat{y} \in D_d) - min(\hat{y} \in D_d)}{\sum_{d=1}^{n} max(y \in D_d) - min(y \in D_d)}$
- $max\_error = \frac{\sum_{h=1}^{24} \#maximum_h(Y) - \#maximum_h(\hat{Y})}{2n}$
- $sign\_cor = \frac{\sum_{h=1}^{24} (\sum_{i \in N_h} \mathbb{1}(y_i > y_{i+1}) - \sum_{i \in N_h} \mathbb{1}(\hat{y}_i > \hat{y}_{i+1}))}{24n}$

Comments 2/2

- n: Number of days
- D_d: Day d
- $maximum_h(Y)$: Returns the days where the maximum of the day is at hour h
- $N_h = \{h + 24z : z \in 1, \ldots, n\}$

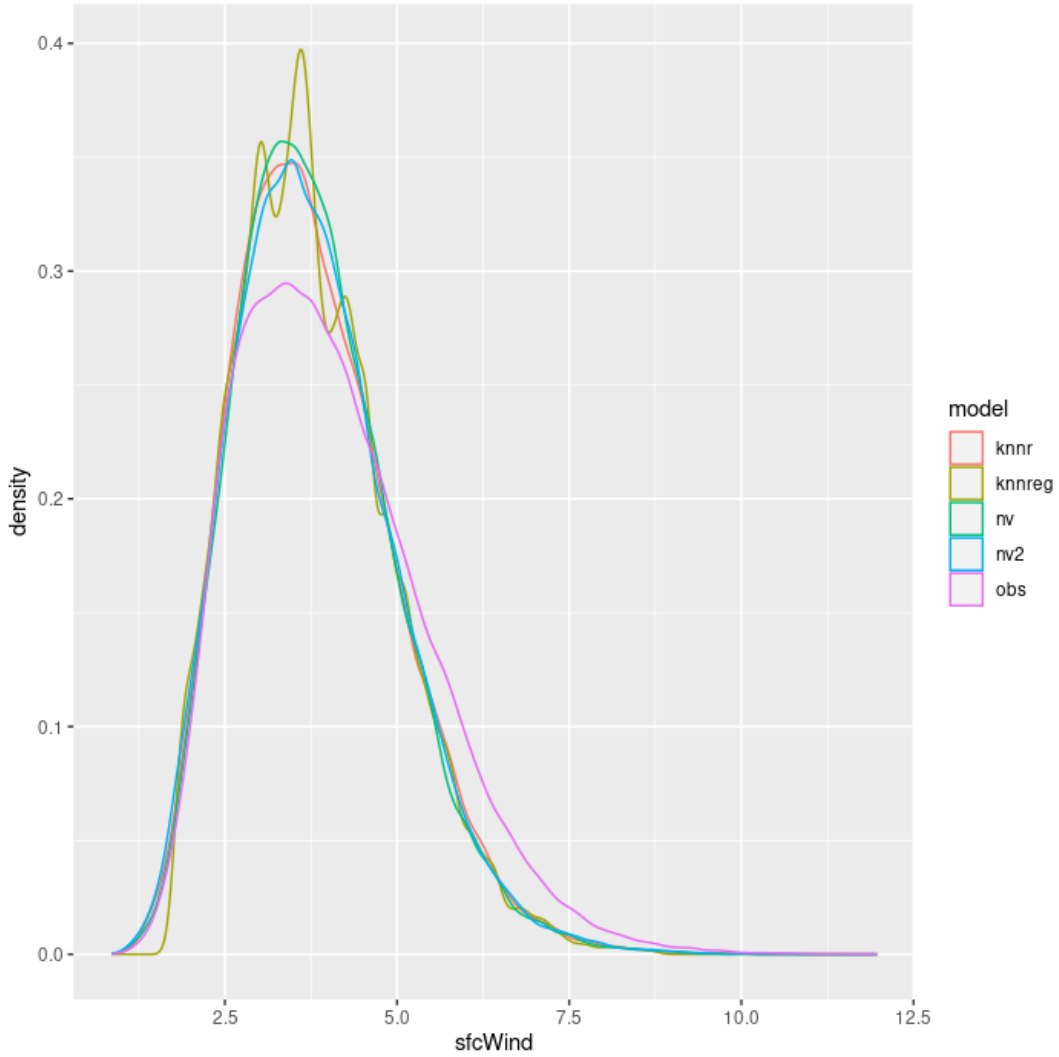
How Often Peaks Hit Hourly



Comments

- On this aspect the worst model is the naive model since practically predict all the peaks in the same hour.
  - The naive 2.0 is a improvement of the naive model on this aspect, but seems that is not good enough.
- The knnr is the one with best performance. Besides some differences predict peaks at different hours.

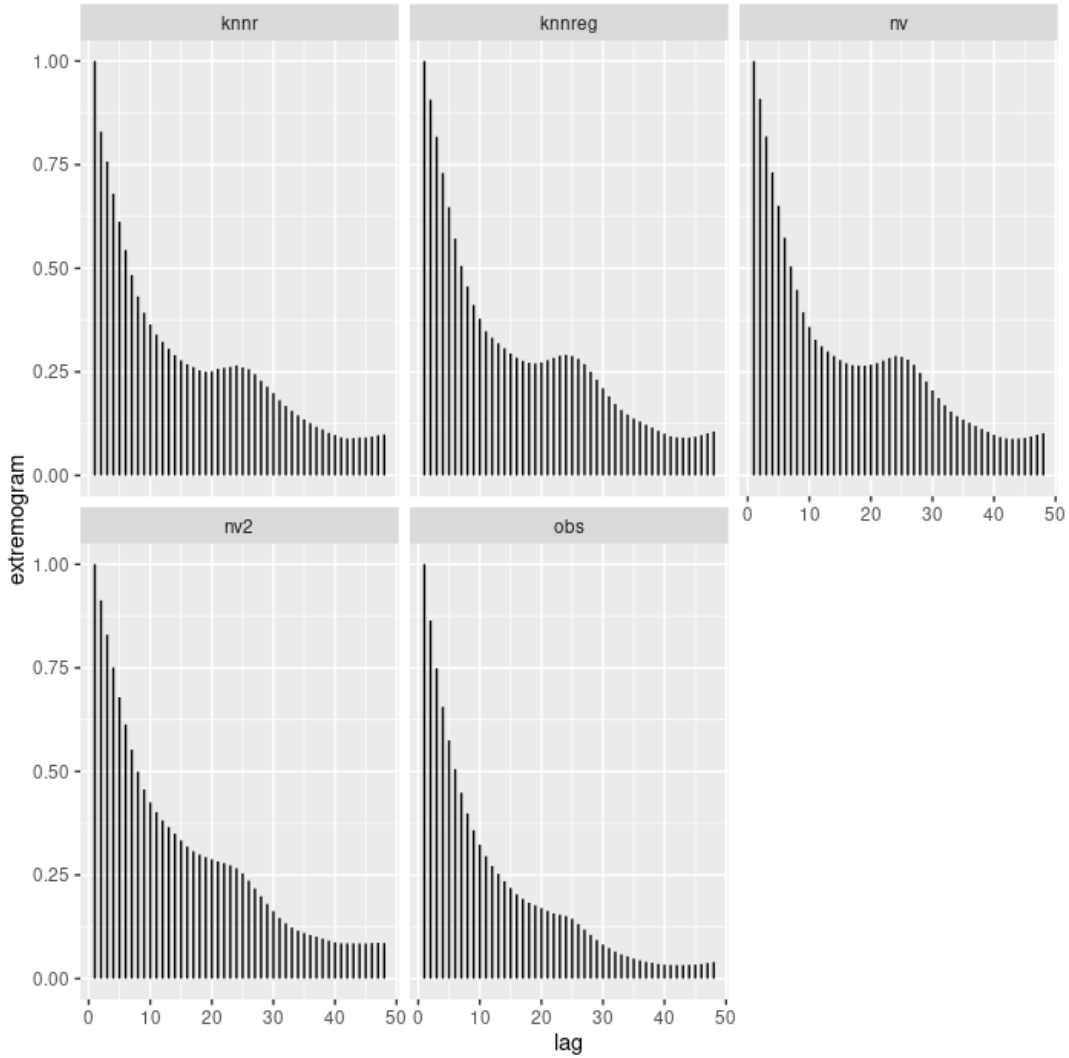
Densities


0.4
0.3
0.2
0.1
0.0
density
2.5
5.0
7.5
10.0
12.5
sfcWind
model
knnr
knnreg
nv
nv2
obs

Comments

- The downscaled distribution of all the models is more concentrated over the mode than the reanalysis distribution.
- The upper tail of the reanalysis is heavier than the downscaled distribution of all the models.
- The knn regression model has a multimodal distribution that is anything like the reanalysis distribution, also gives a near zero probability to the smallest values.

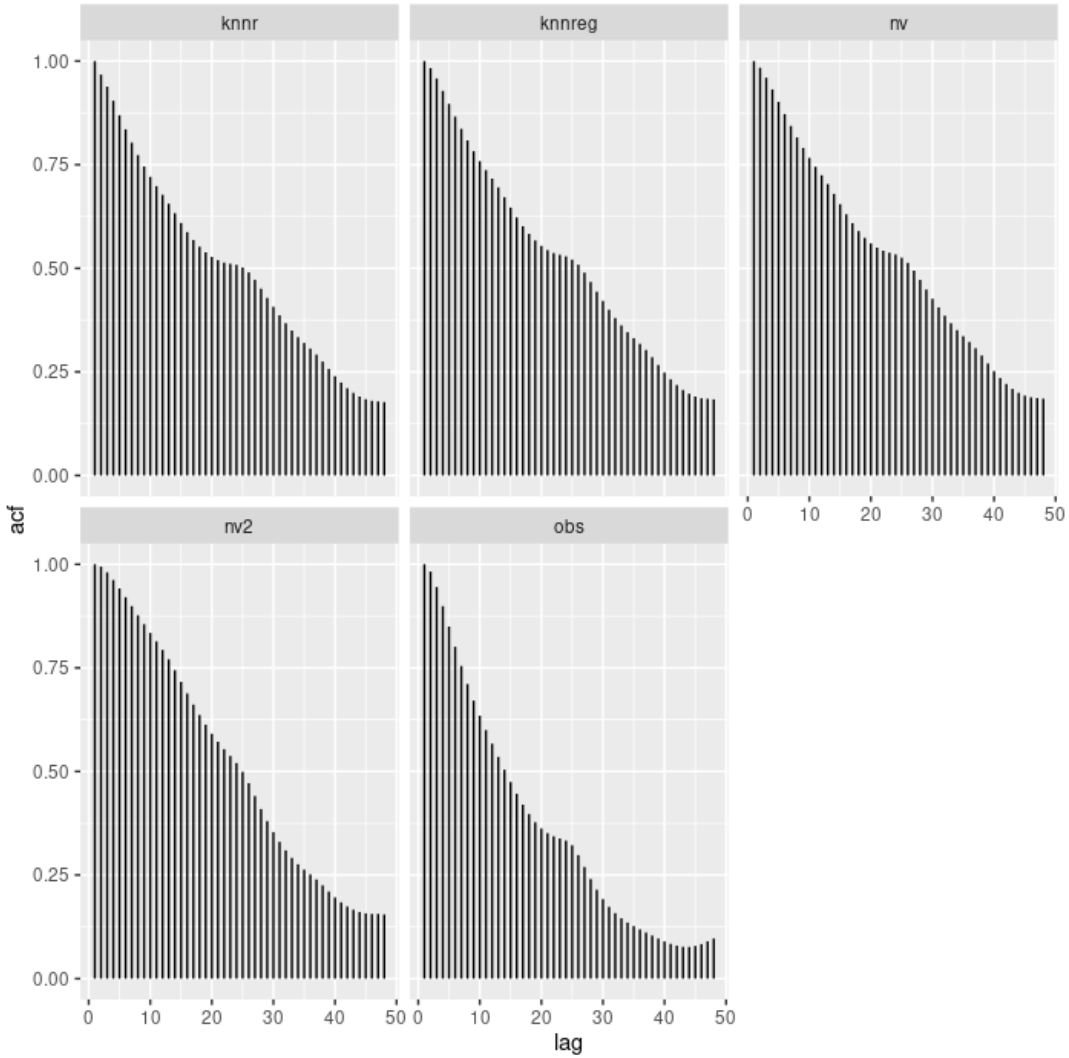
Extremograms



Comments

- In all the models we have that the likelihood of an extreme value appearing with a large lag is consistently overestimated i.e. all the extremograms had a slower decay than the extremogram of the reanalysis.
- In every model, the extremogram shows a rise around the lag 24, indicating that when an extreme value occurs, the next day is more likely to also experience an extreme value, in comparison with the reanalysis.

ACF


knnr
knnreg
nv
nv2
obs
acf
lag
1.00
0.75
0.50
0.25
0.00
0
10
20
30
40
50

Comments

- The reanalysis acf plot has a considerable steepest decay
- An interest result that I don't here is the acf by hour, there we can see that for the hour that are different of the 0 hour the acf is overestimated and for the 0 hour the acf is underestimated.