Nonhomogeneous Boosting for Predictor Selection in Ensemble Postprocessing

Jakob W. Messner, Georg J. Mayr, and Achim Zeileis

Electrical Engineering, Technical University of Denmark (DTU)

Weather forecasts

Numerical Weather Prediction (NWP)

- Observations $\rightarrow$ estimate current atmospheric state.
- Simulate atmospheric processes with numerical models.

$\Rightarrow$ Compute future weather

Weather forecasts

Numerical Weather Prediction (NWP)

- Observations $\rightarrow$ estimate current atmospheric state.
- Simulate atmospheric processes with numerical models.

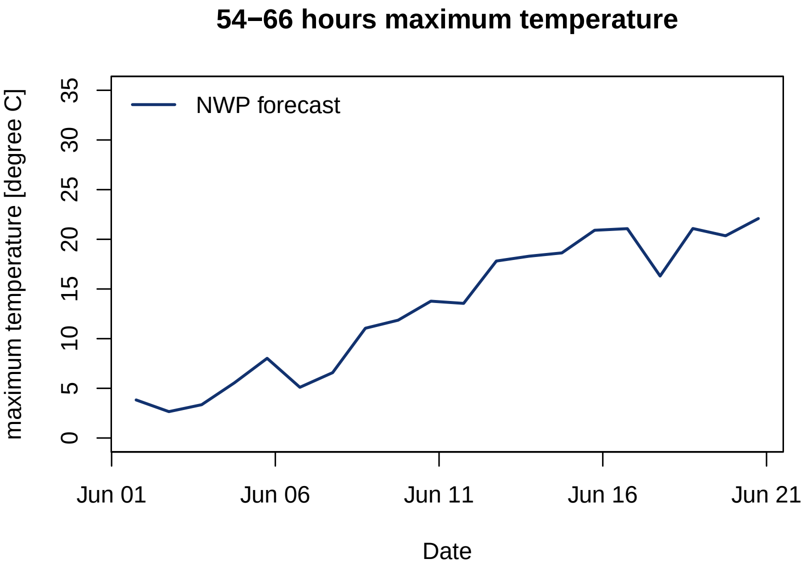
$\Rightarrow$ Compute future weather

Problems:

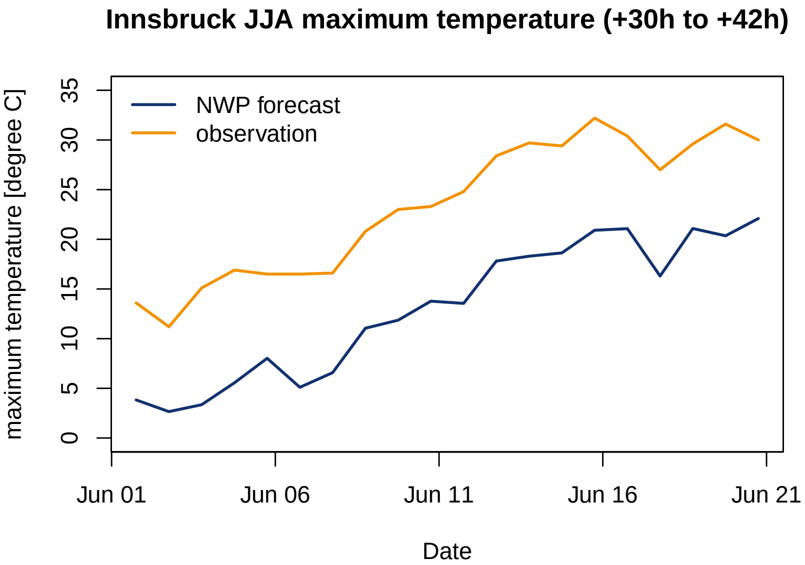
- Few observations
- Observation errors
- Not perfectly known atmospheric processes
- Unresolved processes

$\Rightarrow$ NWP errors

NWP errors

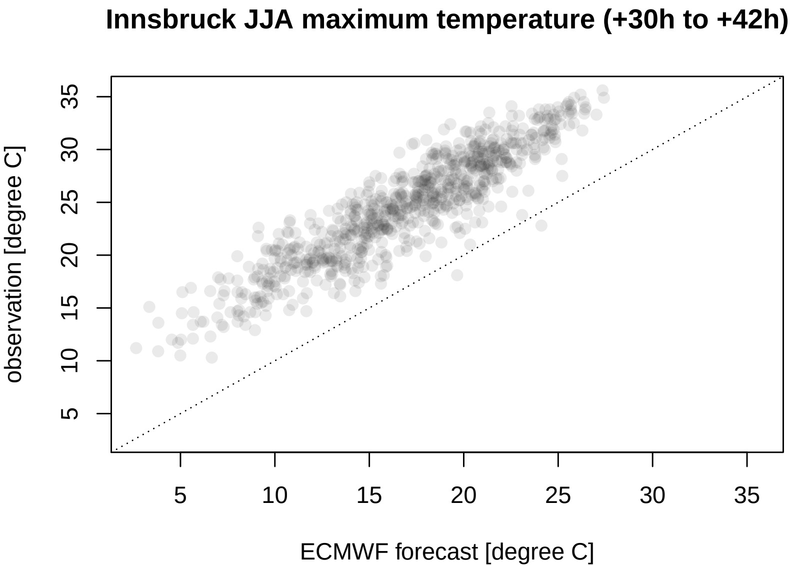

54–66 hours maximum temperature
NWP forecast
maximum temperature [degree C]
0
5
10
15
20
25
30
35
Jun 01
Jun 06
Jun 11
Jun 16
Jun 21
Date

NWP errors

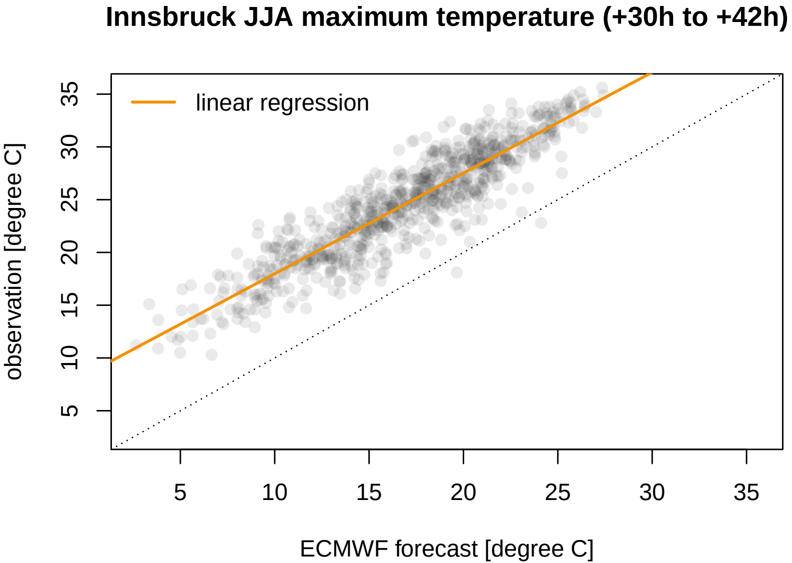


Innsbruck JJA maximum temperature (+30h to +42h)
NWP forecast
observation
maximum temperature [degree C]
0
5
10
15
20
25
30
35
Jun 01
Jun 06
Jun 11
Jun 16
Jun 21
Date

NWP errors


Innsbruck JJA maximum temperature (+30h to +42h)
observation [degree C]
5
10
15
20
25
30
35
ECMWF forecast [degree C]
5
10
15
20
25
30
35

NWP errors


Innsbruck JJA maximum temperature (+30h to +42h)
linear regression
observation [degree C]
ECMWF forecast [degree C]
5
10
15
20
25
30
35
5
10
15
20
25
30
35

Ensemble prediction

NWP error sources:

- Initial conditions
- Model formulations

Ensemble prediction

NWP error sources:

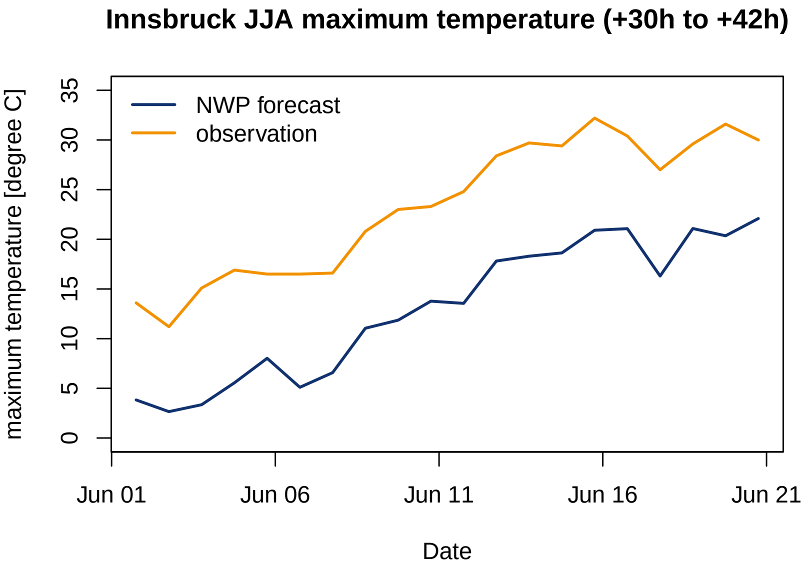
- Initial conditions
- Model formulations

Idea:

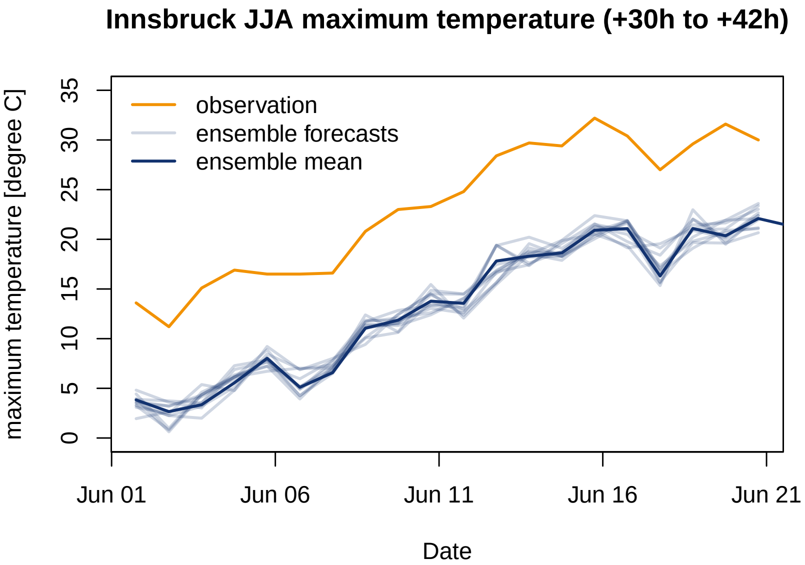
- Perturbed initial conditions
- Different model formulations

$\Rightarrow$ Compute different weather scenarios

NWP errors


Innsbruck JJA maximum temperature (+30h to +42h)
NWP forecast
observation
maximum temperature [degree C]
0
5
10
15
20
25
30
35
Jun 01
Jun 06
Jun 11
Jun 16
Jun 21
Date

NWP errors



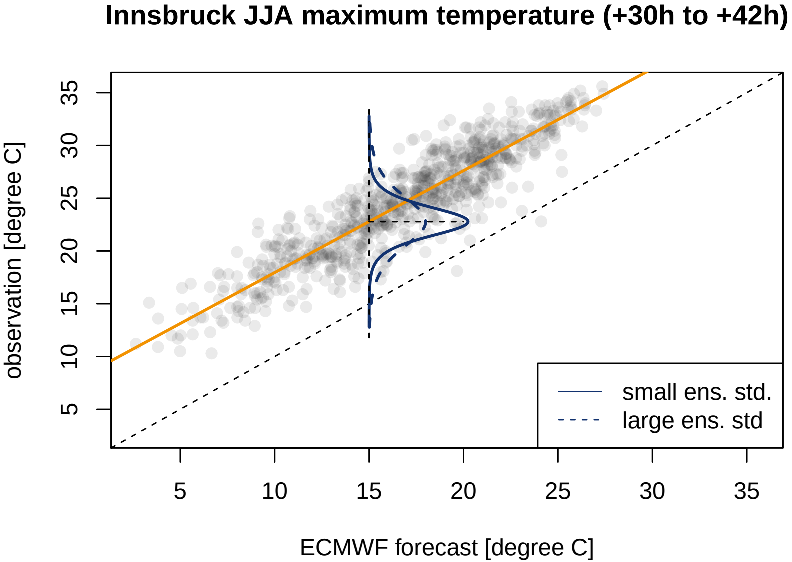
Innsbruck JJA maximum temperature (+30h to +42h)
observation
ensemble forecasts
ensemble mean
maximum temperature [degree C]
0
5
10
15
20
25
30
35
Jun 01
Jun 06
Jun 11
Jun 16
Jun 21
Date

Nonhomogeneous Gaussian regression (NGR)

$$\begin{aligned} y &\sim \mathsf{N}(\mu, \sigma) \\ \mu &= \beta_0 + \beta_1 x \\ \log(\sigma) &= \gamma_0 + \gamma_1 z \end{aligned}$$

y	response (e.g., temperature)
x	ensemble mean (e.g., of temperature ensemble)
z	ensemble standard deviation
$\beta_0, \beta_1, \gamma_0, \gamma_1$	regression coefficients

Nonhomogeneous Gaussian regression (NGR)



Nonhomogeneous Gaussian regression (NGR)

Inputs:

- deterministic MOS: common to use multiple input variables
- NGR: usually only ensemble forecasts of forecast variable (e.g. maximum temperature)
- other potential variables:
  - ensemble forecasts of other variables (pressure, cloud cover, ...)
  - current observations
  - ensemble or deterministic forecasts from other centers
  - transformations or interactions
  - ...

Nonhomogeneous Gaussian regression (NGR)

$$
\begin{aligned}
y &\sim \mathrm{N}(\mu, \sigma) \\
\mu &= \beta_0 + \beta_1 x_1 + \beta_1 x_2 + \ldots + \epsilon = \mathbf{x}^\top \beta \\
\log(\sigma) &= \gamma_0 + \gamma_1 z_1 + \gamma_2 z_2 + \ldots = \mathbf{z}^\top \gamma
\end{aligned}
$$

y	response (e.g., temperature)
$\mathbf{x}$	inputs for location (e.g., different ensemble means and standard deviations, current observations, etc.)
$\mathbf{z}$	inputs for scale (e.g., different ensemble means and standard deviations, current observations, etc.)
β, γ	regression coefficients

Problem: How to select variables in **x** and **z**.

Nonhomogeneous boosting

$$\begin{aligned} y &\sim \mathsf{N}(\mu, \sigma) \\ \mu &= \mathbf{x}^\top \beta \\ \sigma &= \mathbf{z}^\top \gamma \end{aligned}$$

Maximum likelihood estimation:

$$L = \sum \log \left[\frac{1}{\sigma} \Phi \left(\frac{y - \mu}{\sigma} \right) \right]$$

Nonhomogeneous boosting

$$\begin{aligned} y &\sim \mathsf{N}(\mu, \sigma) \\ \mu &= \mathbf{x}^\top \beta \\ \sigma &= \mathbf{z}^\top \gamma \end{aligned}$$

Maximum likelihood estimation:

$$L = \sum \log \left[\frac{1}{\sigma} \Phi \left(\frac{y - \mu}{\sigma} \right) \right]$$

Gradient boosting:

- initialize all coefficients with zero
- in each iteration slightly **update only the one coefficient that improves the current fit most**

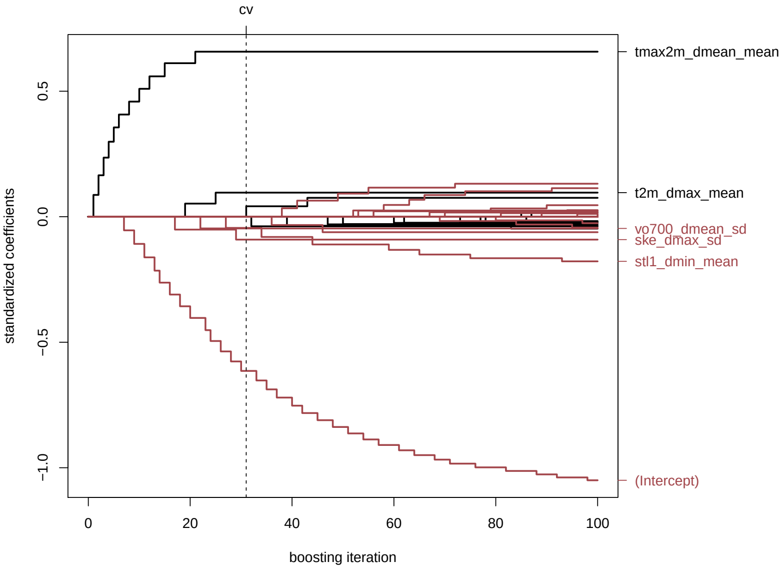
$\rightarrow$ if not run until convergence, **only important inputs have non-zero coefficients**

- select optimum stopping iteration by cross validation

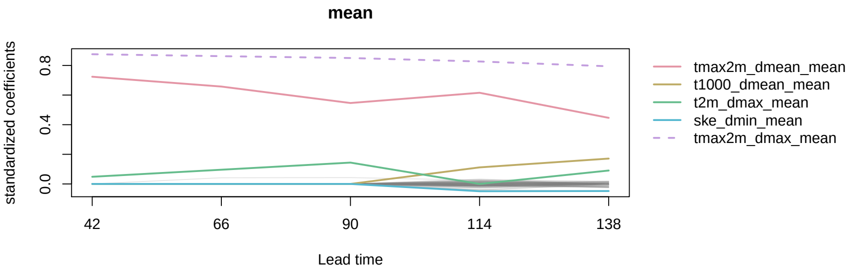
Nonhomogeneous boosting

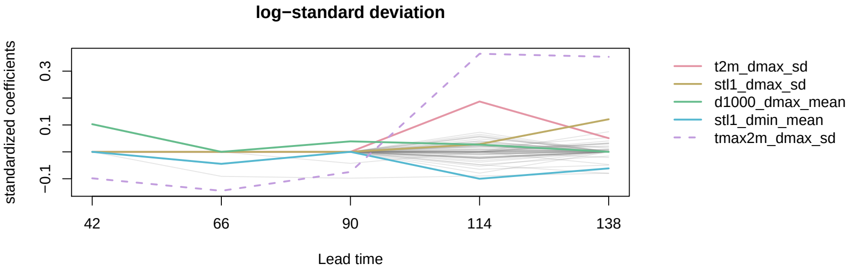
1. initialize $\beta = \mathbf{0}$, $\gamma = \mathbf{0}$
2. $\mu = \mathbf{x}^\top \beta$, $\sigma = \mathbf{z}^\top \gamma$
3. find predictor x_j most correlated with $-\partial L / \partial \mu$ and
   z_k most correlated with $-\partial L / \partial \sigma$
4. update $\beta_j^* \leftarrow \beta_j + \nu \text{cor}(x_j, \partial L / \partial \mu))$ and
   $\gamma_k^* \leftarrow \gamma_k + \nu \text{cor}(z_k, \partial L / \partial \sigma))$ with $0 < \nu < 1$
5. use only update with best likelihood:
   if $L(\mu^*, \sigma) > L(\mu, \sigma^*)$ set $\beta = \beta^*$ else $\gamma = \gamma^*$
6. repeat step 3. to 6. until predefined stopping iteration is reached

Wien 66 hours maximum temperatures

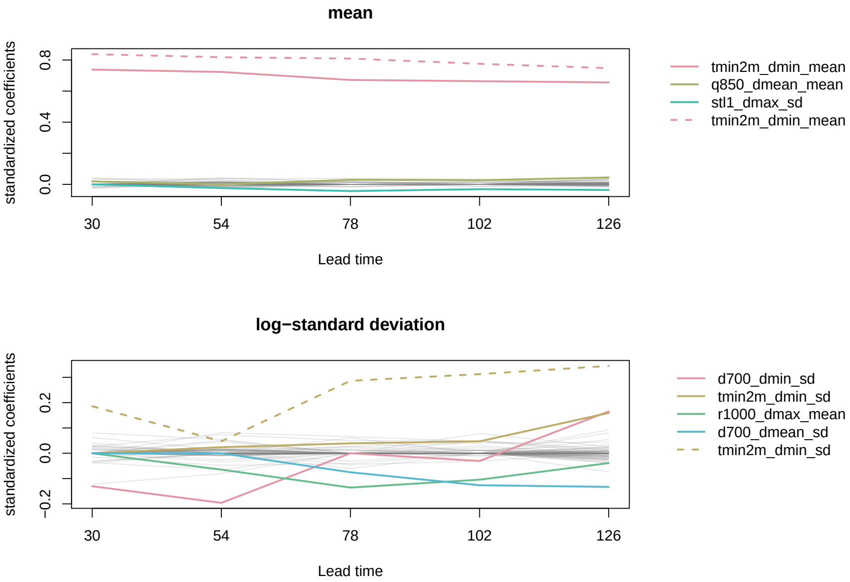


Wien maximum temperature coefficients

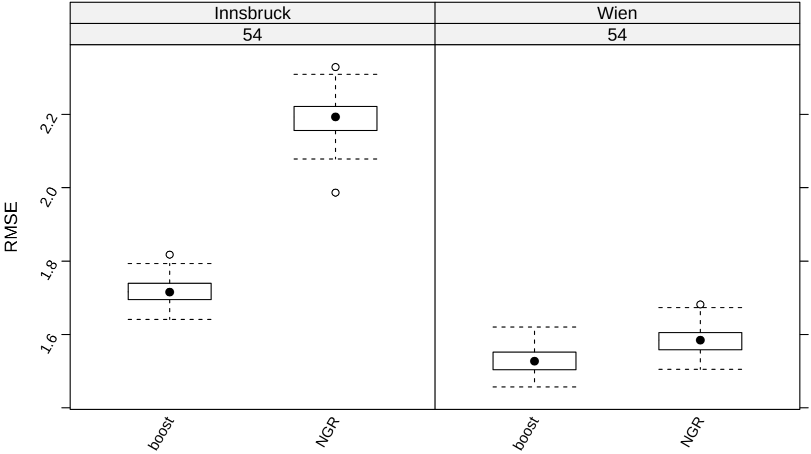




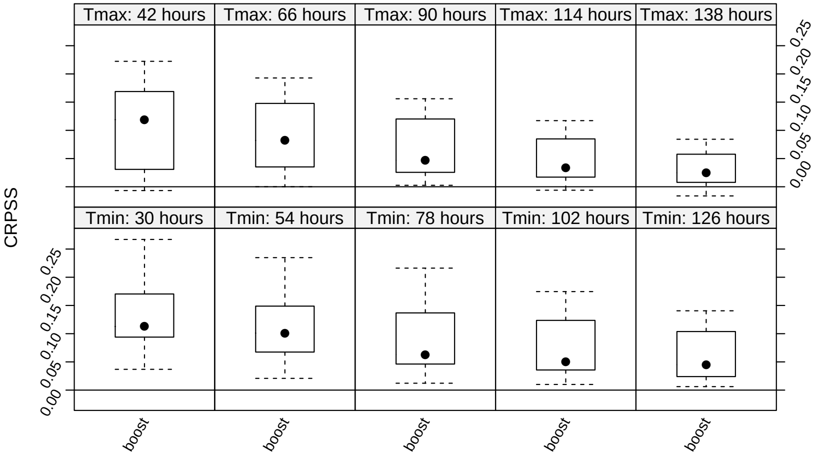
Wien minimum temperature coefficients



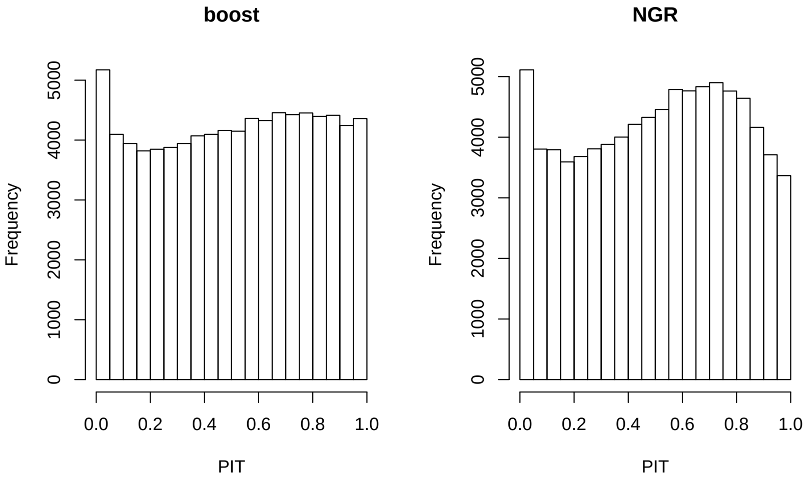
+42 to +54 hours minimum Temperature RMSE



CRPS skill score



PIT histogram



Summary

Nonhomogeneous boosting:

- efficient variable selection
- clearly improved forecast performance compared to common NGR

Summary

Nonhomogeneous boosting:

- efficient variable selection
- clearly improved forecast performance compared to common NGR

References:

Messner, J. W., G. J. Mayr, and A. Zeileis, 2017: Nonhomogeneous boosting for predictor selection in ensemble postprocessing. *Monthly Weather Review*, **145 (1)**, 137–147.