UNIVERSITY OF OSLO

Chapter 9
Multiple regression

Kolbjørn Engeland

koe@nve.no

The regression equation

- $Y = a + b_1X1 + b_2X2 + \cdots, +b_kXk + e$
- Standard assumption: $e \sim N(0, \sigma^2)$
- *X* is called «indenpent variable», «explanatory variable» , «regressor», «covariates», «predictor»
- Y is called «dependent variable» , «response», «predictant»
- No interaction term in this model!

Why multiple regression?

- Analyze
  - Understand hydrology
    - Linear trends
    - Catchment properties controlling hydrology
  - NB: Correlation does not imply causality
- Predict
  - In time
  - In space

Matrix formulation

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad \text{for } i = 1, \ldots, n$$

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_1 + \epsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_2 + \epsilon_2 \\ &\vdots \\ y_n &= \beta_0 + \beta_1 x_n + \epsilon_n \end{aligned}$$

$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}}_{Y} = \underbrace{\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}}_{X} \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}}_{\beta} + \underbrace{\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}}_{\varepsilon}$$

$$Y = X\beta + \varepsilon$$

UNIVERSITY OF OSLO

Alternative formulation

- The residual is then

$$\epsilon = \boldsymbol{y} - \boldsymbol{X\beta}$$

- And the squared residual is:

$$\epsilon'\epsilon = (\boldsymbol{X\beta} - \boldsymbol{y})'(\boldsymbol{y} - \boldsymbol{X\beta})$$

- The derivative is:

$$-2\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X\beta}) = 0$$

$$\boldsymbol{X}'\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{y}$$

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

Standard error of parameters:

- We still use t-test for individual regression coefficients

$$\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}} = \hat{s}^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

Standard errors of regression and predictions:

- Variance for the regression line

$$Var(\hat{\boldsymbol{y}}) = \hat{s}^2 \boldsymbol{X}_*' (\boldsymbol{X}'\boldsymbol{X})^{-1} \boldsymbol{X}_*'$$

- Variance for the prediction:

$$Var(\hat{\boldsymbol{y}}) = \hat{s}^2 \left[1 + \boldsymbol{X}_*' (\boldsymbol{X}'\boldsymbol{X})^{-1} \boldsymbol{X}_*'\right]$$

Categorical variables

- Independent variables might be categorical
  - Equivalent to ANOVA

- Then we transfer it into a binary variable

  - 0 or 1

If we have many categories:

many 0 – 1 dummy variables.

Assumptions

- Linearity
- Normality
- Homoscedadisity
- Independence of residuals

UNIVERSITY OF OSLO

Challenge: linearity

- Transform variables
- Box_Cox is flexible and helps in addition for heteroscedadisity
- Use non-linear transformations. (this is non-linear regression)

Challenge: optimal selection of independent variables

- Pool of independent variables
- Could use all, but this might leed to non-robust predictions.
  - Especially for mall data sets and many independent variables
- We need a strategy to select the optimal number of independent variables!

Measuring fit - punishing for freedom

- Adjusted R2

$$R^2_{adjusted} = 1 - (1 - R^2)\left(\frac{n-1}{n-k-1}\right)$$

- $AIC = n * log(RSS/n) + 2k$

- $BIC = n * log(RSS/n) + k * ln(n)$

Measuring fit - punishing for freedom

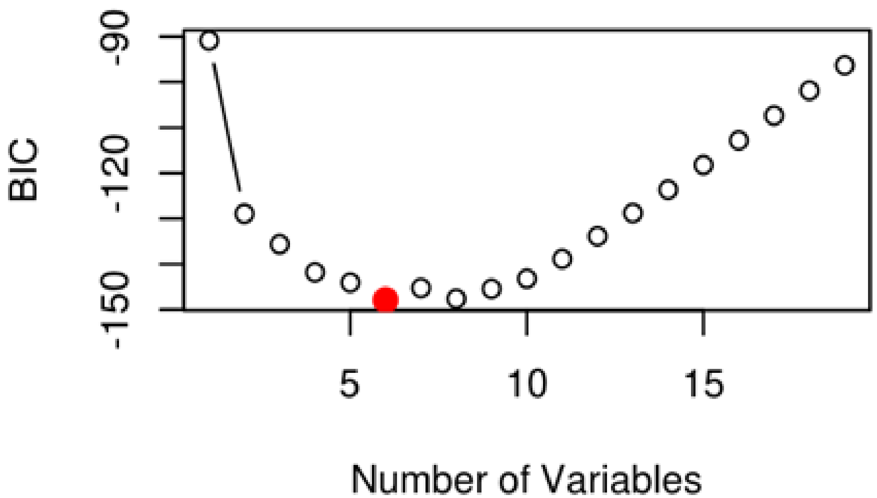
- R2_cv
- As R2, but we use leave-one-out cross-validation to calculate R2 for predicted values.

Strategy for selecting independent variables

- Forward selection
- Backward elemination
- Stepwise
- Brute force
- …..

Forward selection

- You select a group of independent variables to be examined.
- Evaluate the fir the first covariate, one by one using a criterion
- Add the covariate giving the best performance.
- Evaluate the second covariate among the remaining
- Add only if the performance increase



Backward elemination

- You start with all covariates in the model
- Eliminate the covariate leading to an increase in performance, one by one.
- Stops when the performance it at the maximum

Stepwise regression

- This works like forward regression
- But evaluate at each stage, the possibility that a variable entered at a previous stage might now be eliminated because of additional variables now in the model that were not in the model when this variable was selected.

What if assumtions fails?

- Somewhat biased estimates

- Statistical inference and hypothesis testing fails
  - Might be to confident in the results

- Unreliable predictions.
  - The predictions is the most sensitive to model assumptions.

How to evaluate assumtions?

- Linearity: scatter plot
- Normality: Histogram, qq-plot, KS test
- Independence: auto-correlation
- Homoscedadisity. Scatter plot of residuals

Regression – decomposition and modelling of residual variance

WLS and GLS Regression

Observation errors: $\boldsymbol{\Sigma}_\varepsilon$

Model errors: $\boldsymbol{\Sigma}_\delta = \boldsymbol{\delta}^2\boldsymbol{I}$

Total error: $\boldsymbol{\Sigma}_\epsilon = \boldsymbol{\Sigma}_\delta + \boldsymbol{\Sigma}_\varepsilon = \delta^2\boldsymbol{I} + \boldsymbol{\Sigma}_\varepsilon$

Regression coefficients: $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{\Sigma}_\epsilon^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}_\epsilon^{-1}\boldsymbol{y}$

Prediction variance: $Var(\widehat{\boldsymbol{y}}) = \delta^2 + \boldsymbol{X}_*'(\boldsymbol{X}'\boldsymbol{\Sigma}_\epsilon^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}_*$

Regression – decomposition and modelling of residual variance

WLS and GLS Regression

Challenge: estimate $\boldsymbol{\delta^2}$

Sollution: iterative approach

1: Estimate $\widehat{\boldsymbol{\beta}}$ $\qquad \widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{\Sigma}_{\epsilon}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}_{\epsilon}^{-1}\boldsymbol{y}$

2: Estimate $\boldsymbol{\delta^2}$ $\qquad \left(\boldsymbol{y} - \widehat{\boldsymbol{\beta}}\boldsymbol{X}\right)'(\delta^2\boldsymbol{I} + \boldsymbol{\Sigma}_{\varepsilon})^{-1}\left(\boldsymbol{y} - \widehat{\boldsymbol{\beta}}\boldsymbol{X}\right) = n - k - 1$

Regression – decomposition and modelling of residual variance

WLS: $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$ is diagonal

GLS: $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$ include covariances

Challenge : co-linearity

- Are the independent variables correlated?
- Exact co-linearity: no unique sollution
- High correlation: unstable sollutions
- Select independent variables that have a minimum correlation!
- Use Lasso, ridge or elastic net

Alternative regression approaches

Lasso Regression (Hastie et al, 2009)

- Standardize the independent variables
- Penalty for high values for the regression coeficients
  - The sum of absolute values of the regression coefficients
- Effect: sparse models with fewer coefficients, as some coefficients may become zero and eliminated from the model
- Predictions: smaller estimation variance but a larger bias as compared to OLS

Hastie, T., Tibshirani, R. and Friedman, J.: The elements of statistical learning: prediction, inference and data mining, Springer-Verlag, New York, 2009.

UNIVERSITY OF OSLO

Alternative regression approaches

Ridge Regression (Hoerl and Kennard, 1970, Draper and Smith, 1998)

- Standardize the independent variables
- Penalty is the sum of squared regression coefficients.
- Effect: does not result in an elimination of coefficients, but enables reliable estimates of all parameters.
- It has strong similarities to regression on the principal component of the catchment characteristics.

Hoerl, A. E. and Kennard, R. W.: Ridge regression: Biased estimation for nonorthogonal problems, Technometrics, 12(1), 55–67, 1970.

Draper, N. R. and Smith, H.: Applied regression analysis, John Wiley & Sons., 1998.

Alternative regression approaches

Elastic net regression (Zou and Hastie, 2005)

- Elimination of predictors in Lasso regression is sensitive to the underlying data
- Combines lasso and ridge regression

Zou, H. and Hastie, T. (2005) Regularization and variable selection via the elastic net J. R. Statist. Soc. B (2005) 67, Part 2, pp. 301–320

Variations over linear regression

- Bayesian regression: introduce priors
- Assume other distributions than Normal distribution.

Regression and machine learning

Support vector regression

Regression tree

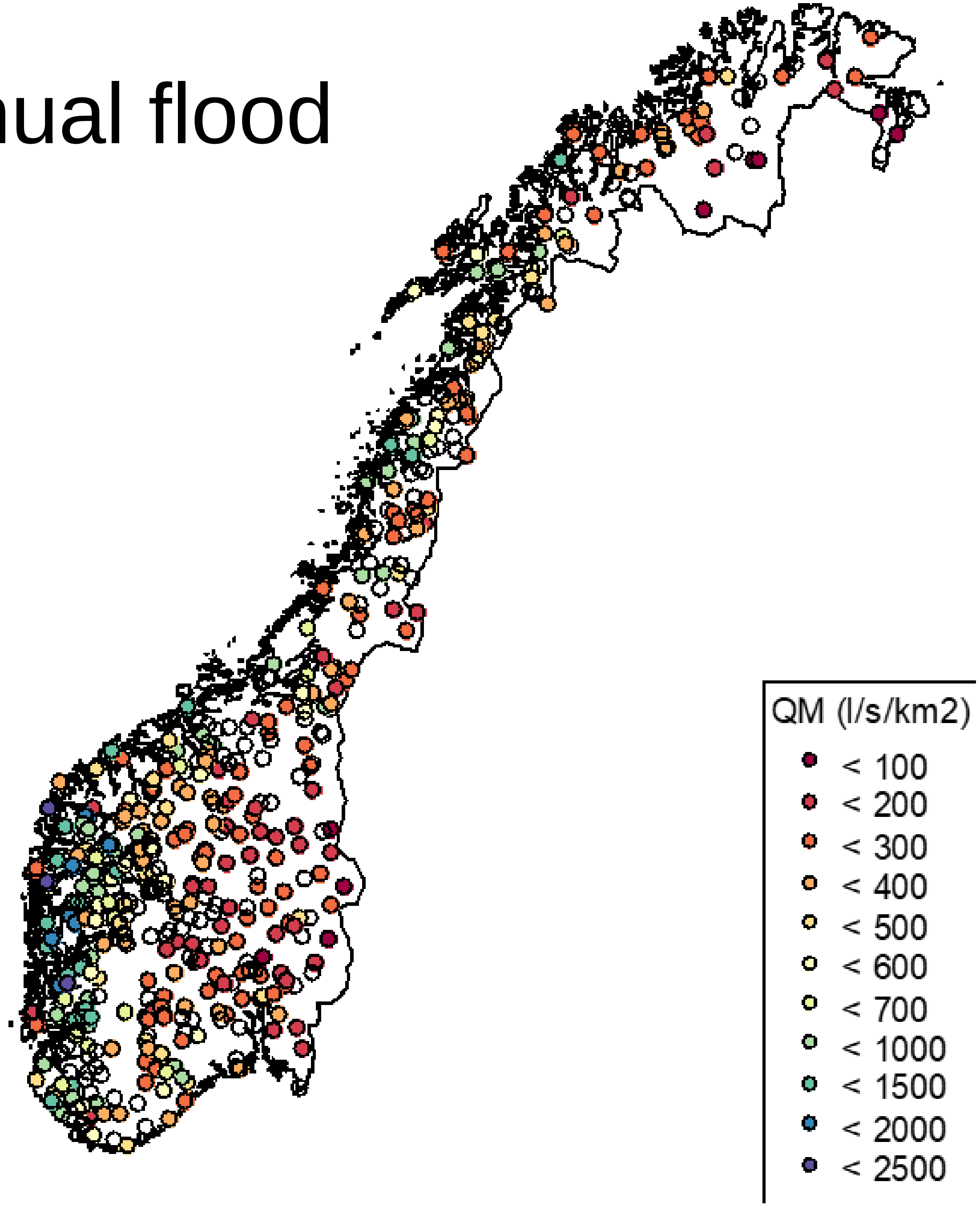
Random Forest Regression

Machine learning library in Python: sklearn

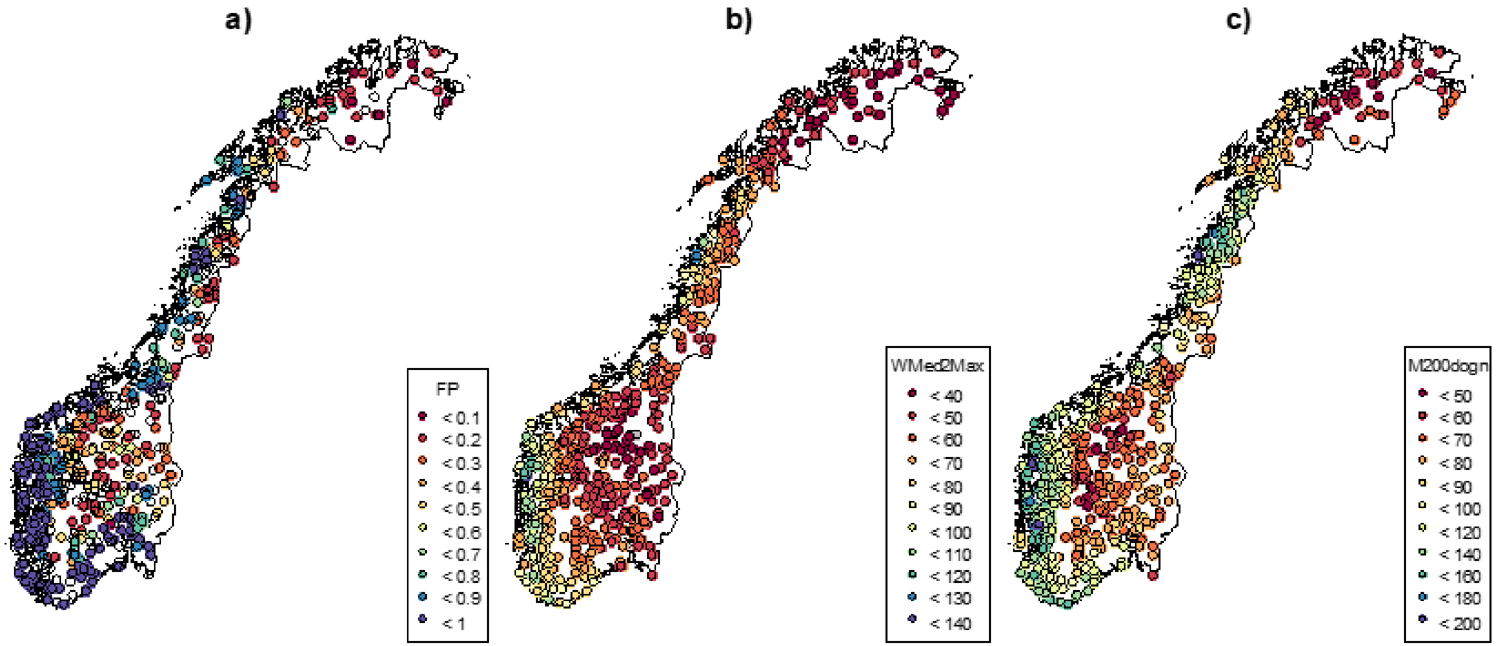
Case study

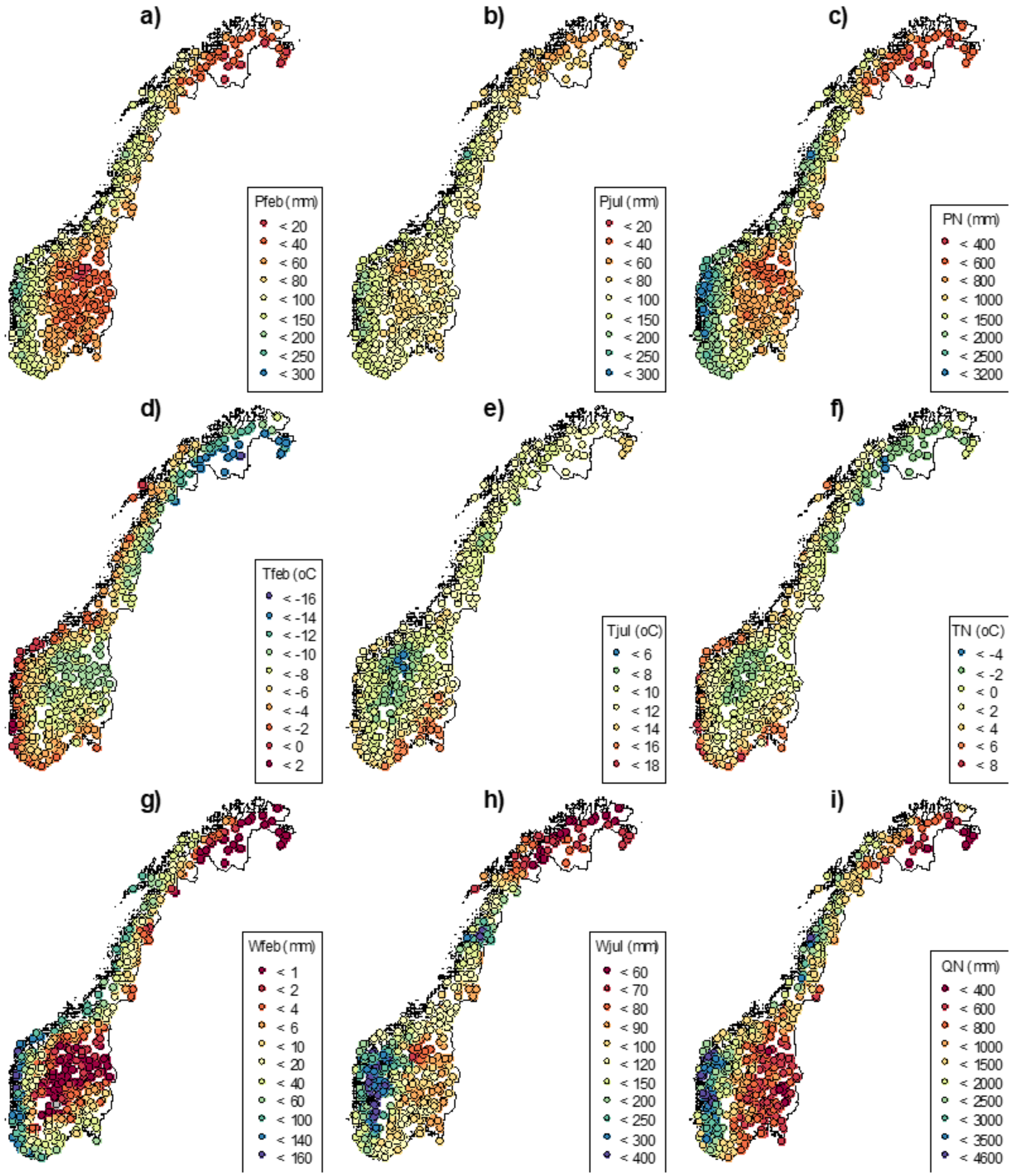
Predict mean annual flow in ungauged basins

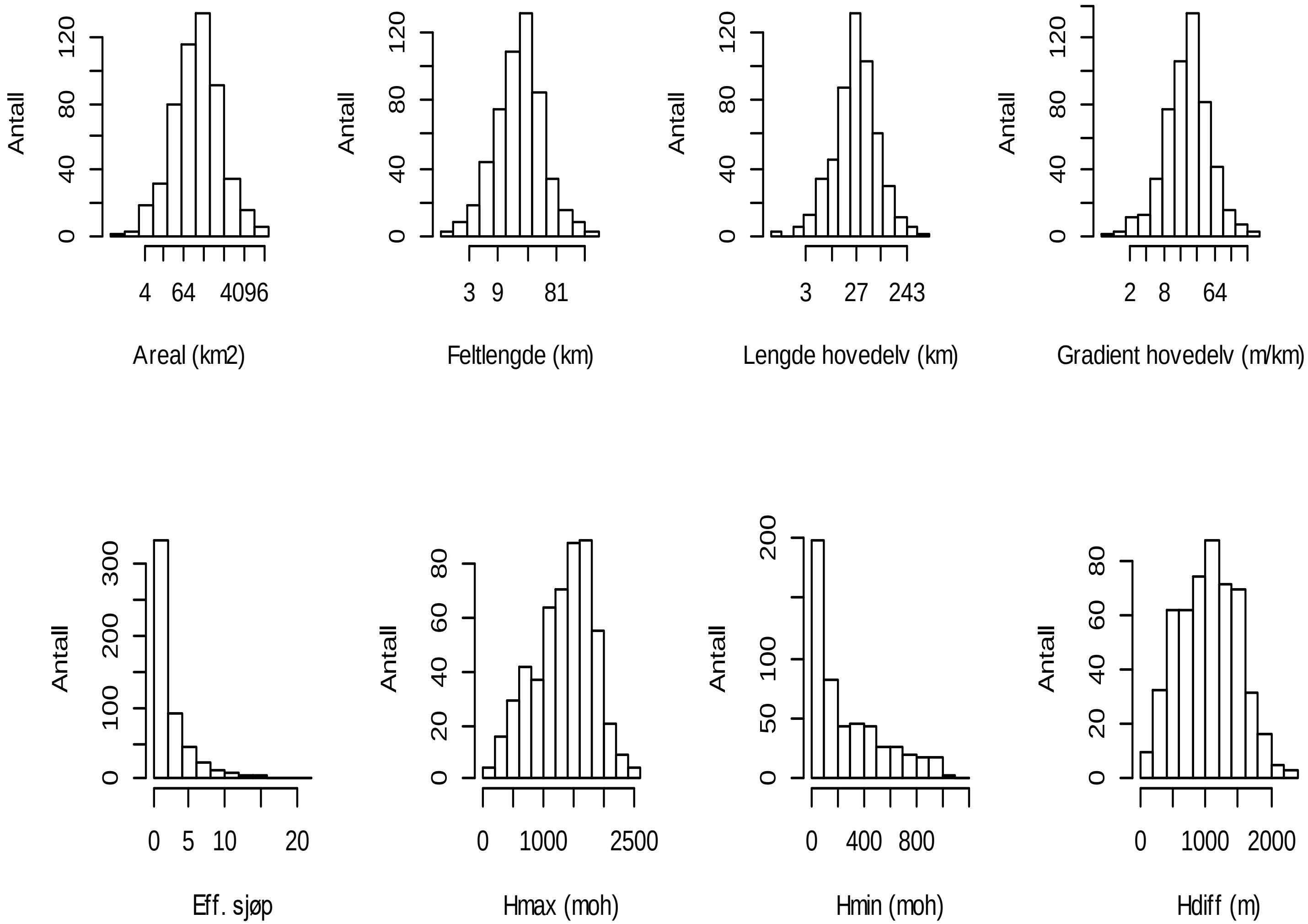
Mean annual flood



Predictors


a)
b)
c)
FP
< 0.1
< 0.2
< 0.3
< 0.4
< 0.5
< 0.6
< 0.7
< 0.8
< 0.9
< 1
WMed2Max
< 40
< 50
< 60
< 70
< 80
< 90
< 100
< 110
< 120
< 130
< 140
M200dogn
< 50
< 60
< 70
< 80
< 90
< 100
< 120
< 140
< 160
< 180
< 200


a)
Pfeb (mm)
< 20
< 40
< 60
< 80
< 100
< 150
< 200
< 250
< 300
b)
Pjul (mm)
< 20
< 40
< 60
< 80
< 100
< 150
< 200
< 250
< 300
c)
PN (mm)
< 400
< 600
< 800
< 1000
< 1500
< 2000
< 2500
< 3200
d)
Tfeb (oC
< -16
< -14
< -12
< -10
< -8
< -6
< -4
< -2
< 0
< 2
e)
Tjul (oC)
< 6
< 8
< 10
< 12
< 14
< 16
< 18
f)
TN (oC)
< -4
< -2
< 0
< 2
< 4
< 6
< 8
g)
Wfeb (mm)
< 1
< 2
< 4
< 6
< 10
< 20
< 40
< 60
< 100
< 140
< 160
h)
Wjul (mm)
< 60
< 70
< 80
< 90
< 100
< 120
< 150
< 200
< 250
< 300
< 400
i)
QN (mm)
< 400
< 600
< 800
< 1000
< 1500
< 2000
< 2500
< 3000
< 3500
< 4600


Antall
Areal (km2)
Feltlengde (km)
Lengde hovedelv (km)
Gradient hovedelv (m/km)
Eff. sjøp
Hmax (moh)
Hmin (moh)
Hdiff (m)

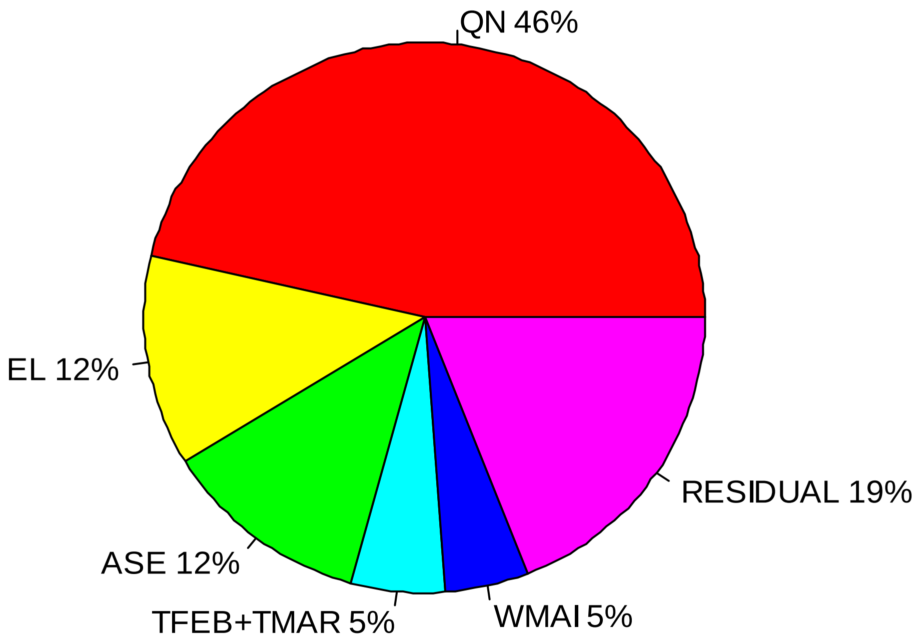
Regression approach

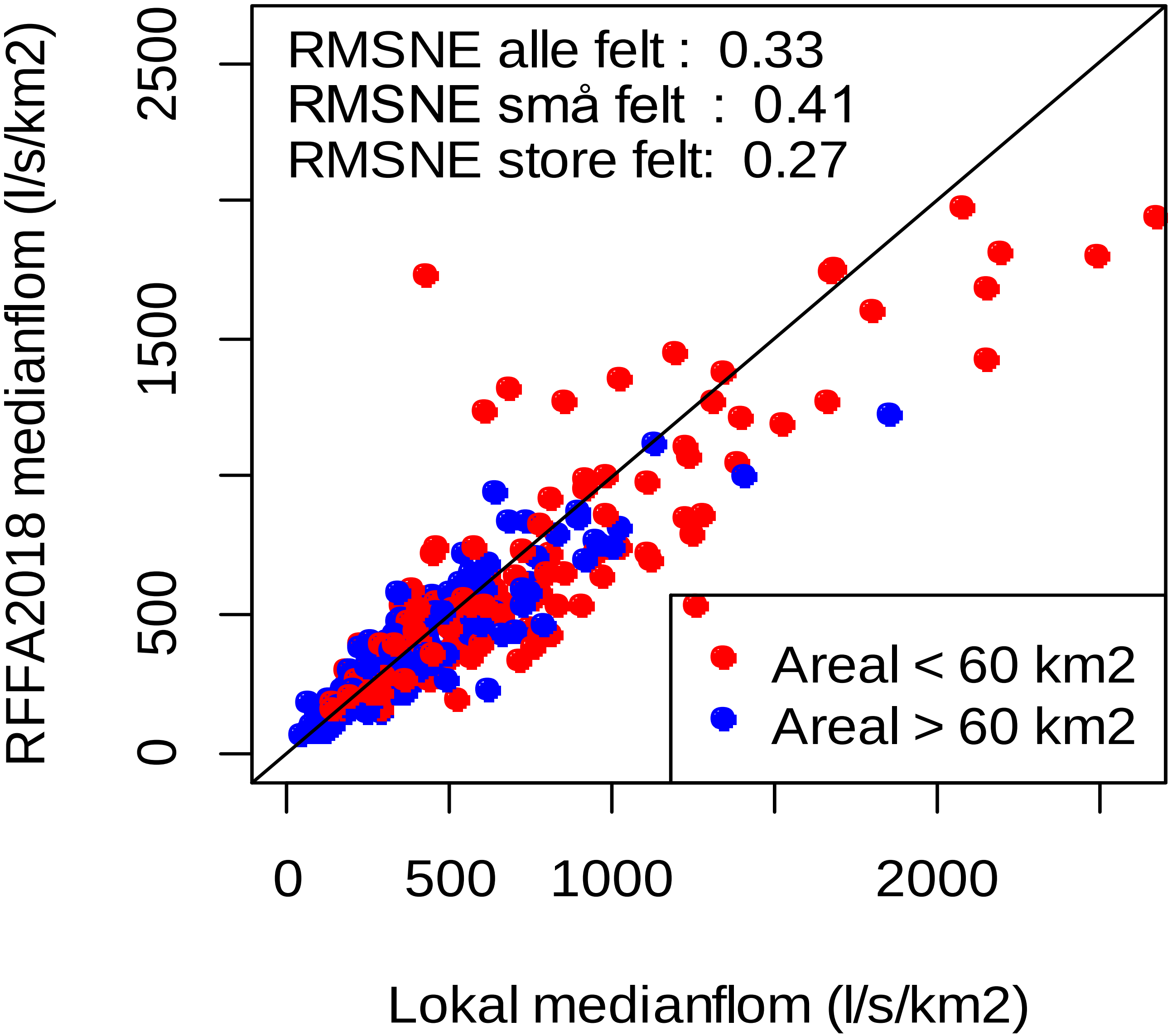
- Log-transform the mean annual flood
- Several transformations for the predictors were tested.
- Flexible approach for assessing required penalty for model complexity
- Dataset divided into three parts
- First set used for selecting predictors and estimating parameters using different penalties
- Second set used for selecting the model (and assosiated penalty) with the smallest RMSE
- Third set used as independent for cross-validation
- Final model: Use the identified predictors and estimate the model once more using all data

Final model

$$q_{ind} = \exp\left[\begin{array}{c} 4.196 + 0.473 * \sqrt[3]{Q_N} - 0.0632 * \sqrt[2]{E_L} - 0.0520 * A_{SE} \\ -0.00751 * {T_{Feb}}^2 - 0.000942 * \ {T_{Mar}}^3 + 0.0376 * \sqrt{W_{Mai}} \end{array}\right]$$

Test sett RMSE var 0.276 (full: 0.262). Dette betyr at med 95% sannsynlighet vil faktisk indeksflom for et felt være innenfor indeksflom-estimat*/1.72.




RMSNE alle felt : 0.33
RMSNE små felt : 0.41
RMSNE store felt: 0.27
Areal < 60 km2
Areal > 60 km2
0
500
1500
2500
0
500
1000
2000
RFFA2018 medianflom (l/s/km2)
Lokal medianflom (l/s/km2)

Excample 1: post-processing

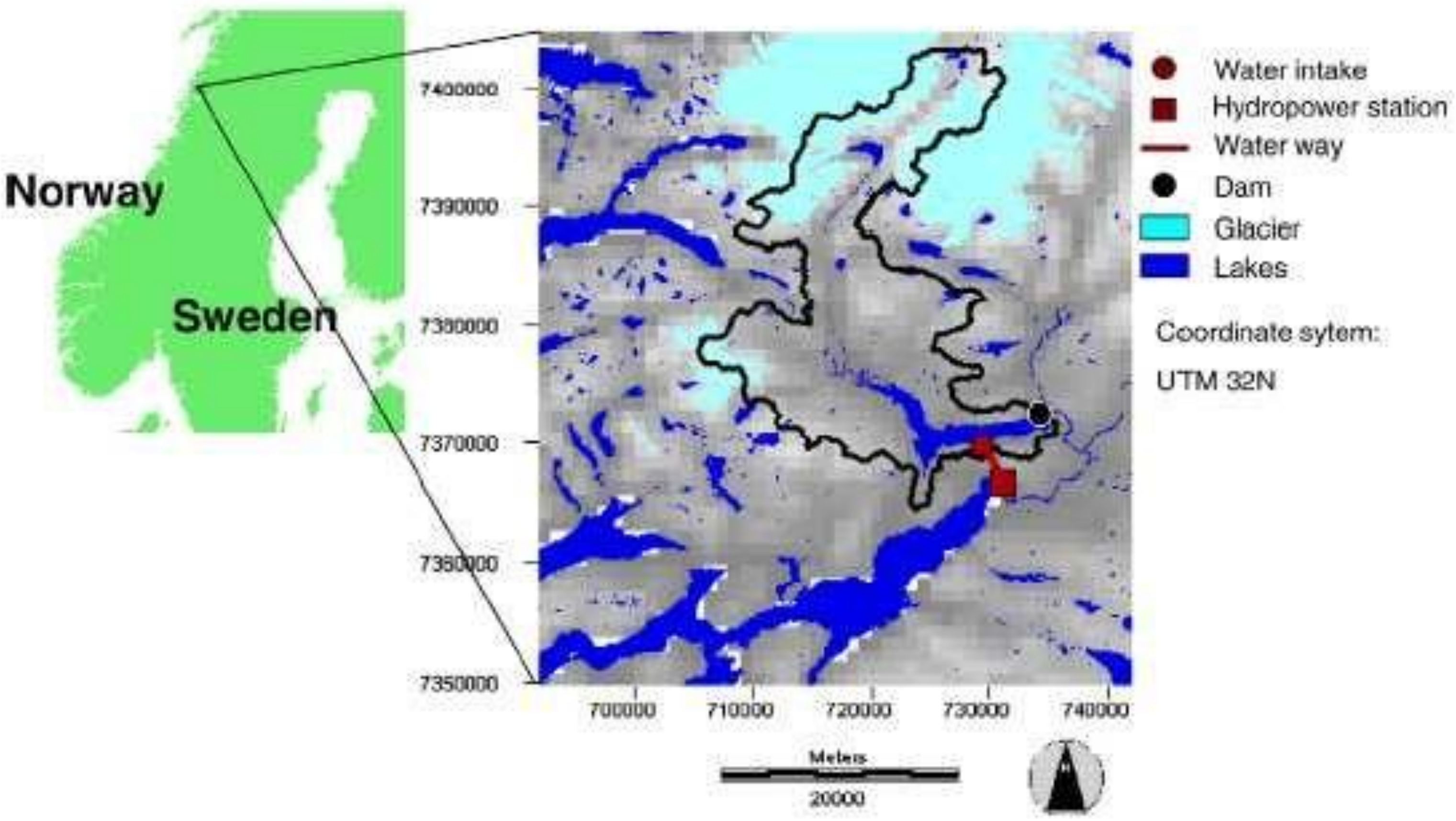


Fig. 1. Map of Langvatn catchment.

1 day forecast

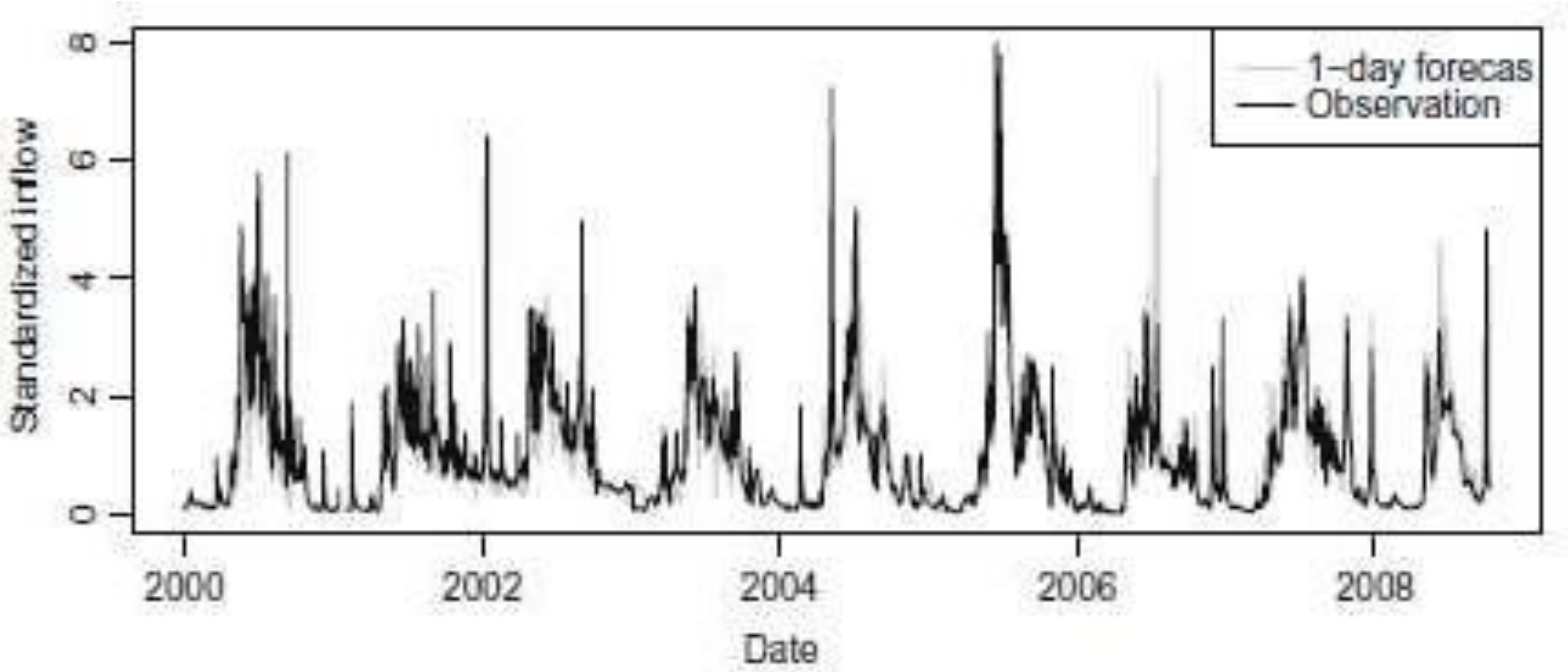
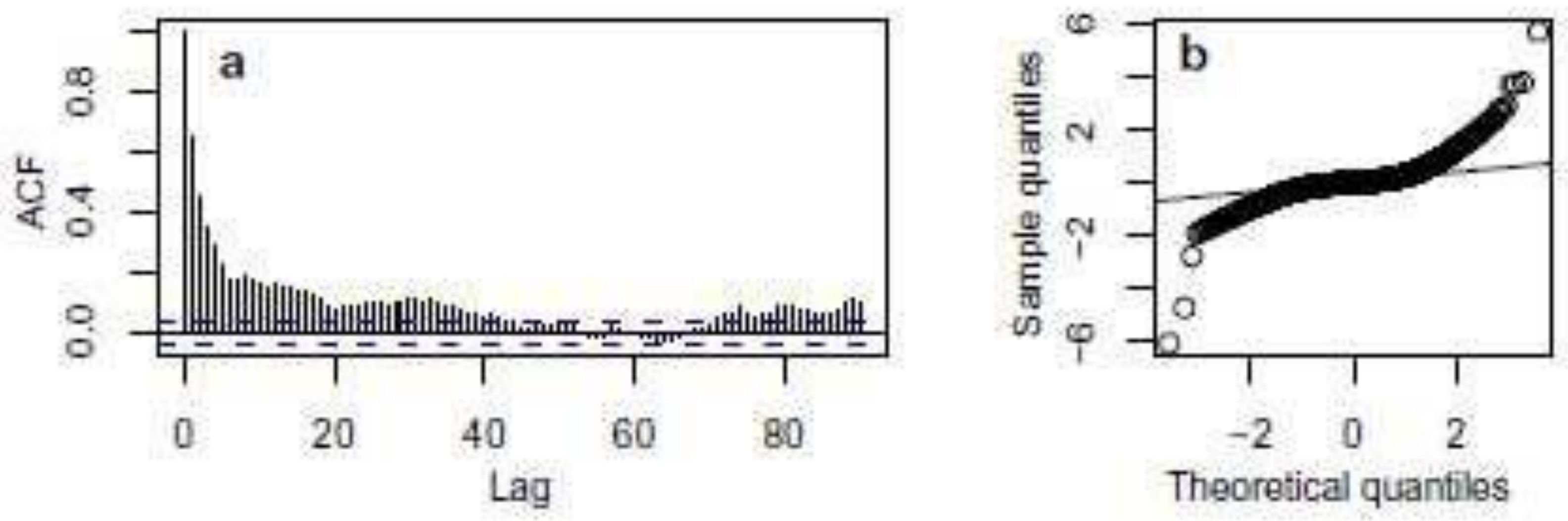
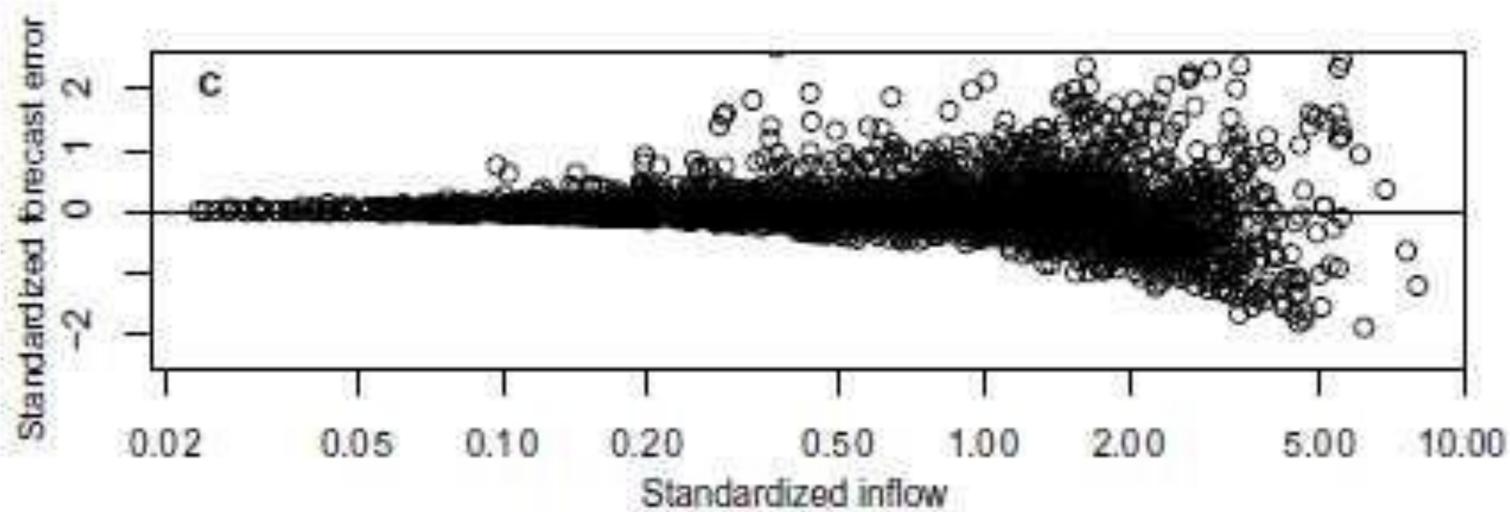


Fig. 2. Observed and 1-day ahead forecasted inflow at Langvatn standardized by the average observed inflow.

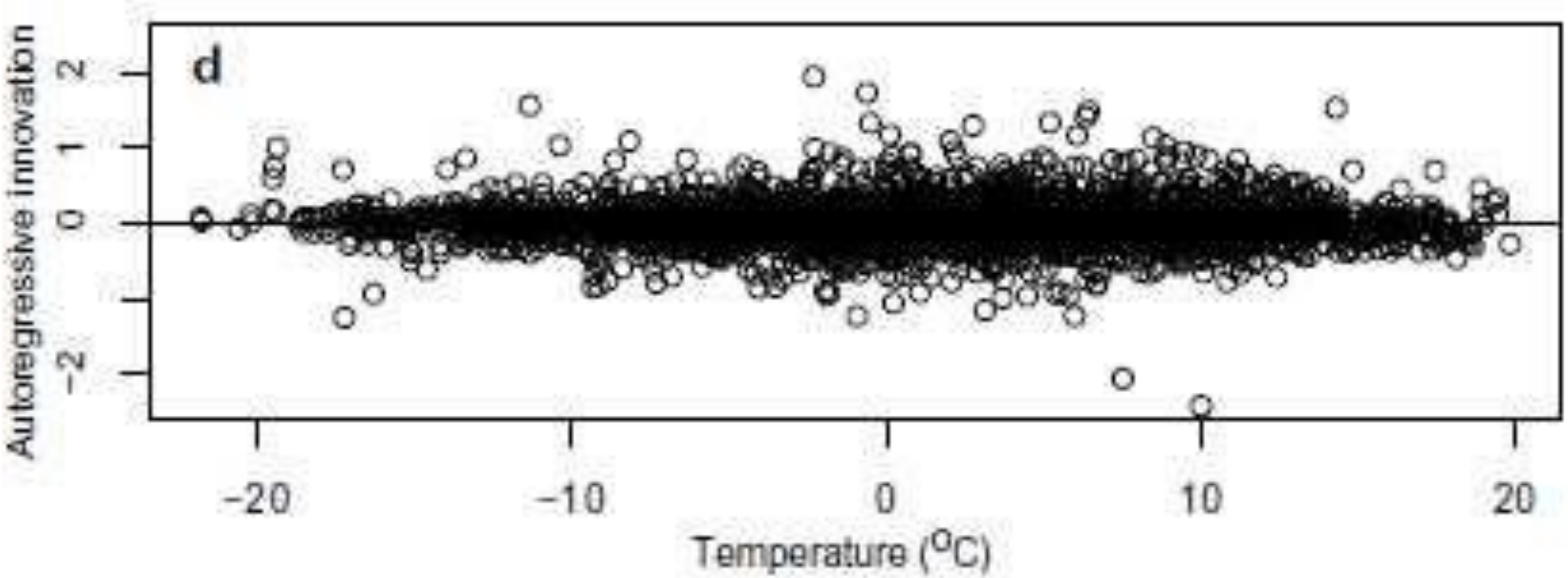
Challenges – independence and distribution



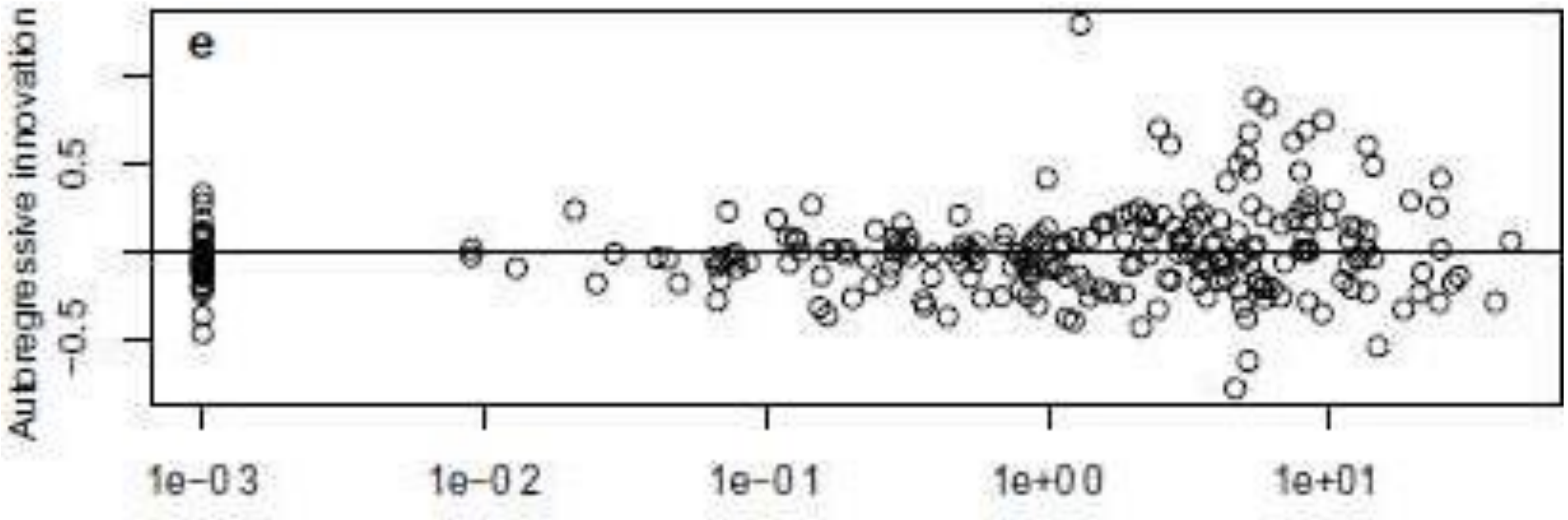
Challenges – constant variance



Challenge: constant variance



Challenge – constant variance



Building model

- Information used:
- Observed precipitation at forecast time.
- Observed temperature at forecast time.
- Forecasted inflow 1-day ahead of forecast time.
- The forecast errors from the previous forecast.
- The season/day of the year.

Building model

- Transformation:

$$q(Q,\lambda) = \begin{cases} \frac{Q^{\lambda}-1}{\lambda} & \lambda \neq 0 \\ \ln(Q) & \lambda = 0 \end{cases}$$

Forecast errors δ for transformed values is then:

$$\delta_t = q_{t,obs} - q_{t,pred}$$

$$\delta_t - (b_{k(t)}) = (a_{k(t)})(\delta_{t-1} - (b_{k(t-1)})) + \varepsilon_t, \varepsilon_t \sim N(0, s_{k(t)})$$

Table 1
Weather classes used in the models for the forecast errors.

Weather class	Temperature (°C)	Precipitation (mm)	Season
1	10.0–20.0	>2.0	Autumn
2	10.0–20.0	>2.0	Spring
3	10.0–20.0	⩽2.0	Autumn
4	10.0–20.0	⩽2.0	Spring
5	5.0–10.0	>2.0	Autumn
6	5.0–10.0	>2.0	Spring
7	5.0–10.0	⩽2.0	Autumn
8	5.0–10.0	⩽2.0	Spring
9	−2.5 to 5.0	>2.0	Autumn
10	−2.5 to 5.0	>2.0	Spring
11	−2.5 to 5.0	⩽2.0	Autumn
12	−2.5 to 5.0	⩽2.0	Spring
13	−10.0 to −2.5	>2.0	All year
14	−10.0 to −2.5	<2.0	All year
15	−30.0 to −10.0	⩾0.0	All year

UNIVERSITY OF OSLO

Table 4
Link between weather classes and the parameters listed in Table 3.

Weather class	Model 1			Model 2		
	b	*s*	*a*	*b*	*s*	*a*
1	b_1	s_1	a_1	b_1	s_1	a_1
2	b_2	s_1	a_2	b_1	s_1	a_2
3	b_3	s_2	a_1	b_2	s_2	a_2
4	b_3	s_2	a_1	b_3	s_3	a_2
5	b_4	s_3	3	b_4	s_4	a_1
6	b_4	s_3	a_2	b_4	s_5	a_2
7	b_3	s_2	a_2	b_2	s_2	a_1
8	b_3	s_1	a_2	b_2	s_1	a_2
9	b_4	s_4	a_1	b_4	s_1	a_2
10	b_1	s_4	a_2	b_4	s_1	a_2
11	b_2	s_1	a_2	b_1	s_3	a_2
12	b_5	s_2	a_2	b_1	s_5	a_1
13	b_2	s_5	a_3	b_1	s_2	a_2
14	b_5	s_6	a_3	b_1	s_2	a_1
15	b_5	s_6	a_3	b_1	s_3	a_1

Table 2
AIC for different versions of Models 1 and 2.

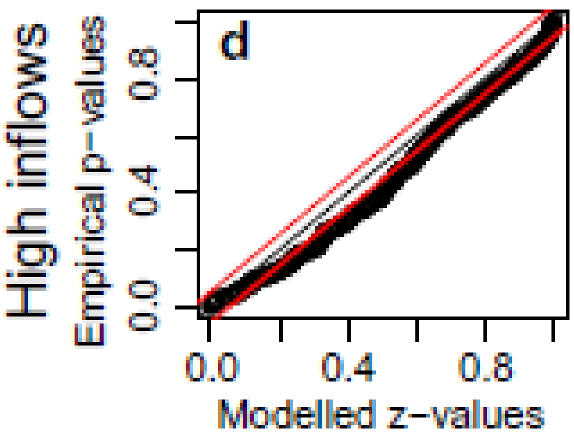
Model version	Model 1		Model 2	
	n	AIC	n	AIC
No weather dependent parameters	3	6974	3	880
b depends on climate	17	7297	17	794
s depends on climate	17	6593	17	627
a depends on climate	17	6953	17	864
b and s depend on climate	31	6524	31	554
b and a depend on climate	31	6865	31	783
s and a depend on climate	31	6566	31	602
All depend on climate	45	6505	45	539
Merged climate classes	14	6451	13	504

UNIVERSITY OF OSLO

Table 3
Estimated parameters for Models 1 and 2. Table 4 shows how parameters and weather classes are linked.

Model 1			Model 2		
Parameter	Estimate	SE	Parameter	Estimate	SE
b_1	0.0879	0.0223	b_1	−0.0018	0.0189
b_2	0.0600	0.0194	b_2	−0.108	0.0226
b_3	−0.0311	0.0190	b_3	−0.0632	0.0293
b_4	0.129	0.0223	b_4	0.0622	0.0218
b_5	0.0279	0.0186	s_1	0.313	0.00745
s_1	0.273	0.00881	s_2	0.213	0.00467
s_2	0.225	0.00564	s_3	0.252	0.00563
s_3	0.376	0.0175	s_4	0.379	0.0229
s_4	0.313	0.00895	s_5	0.444	0.0324
s_5	0.196	0.00755	a_1	0.780	0.0167
s_6	0.172	0.00458	a_2	0.728	0.0155
a_1	0.616	0.0297			
a_2	0.738	0.0186			
a_3	0.833	0.0150			

Evaluation



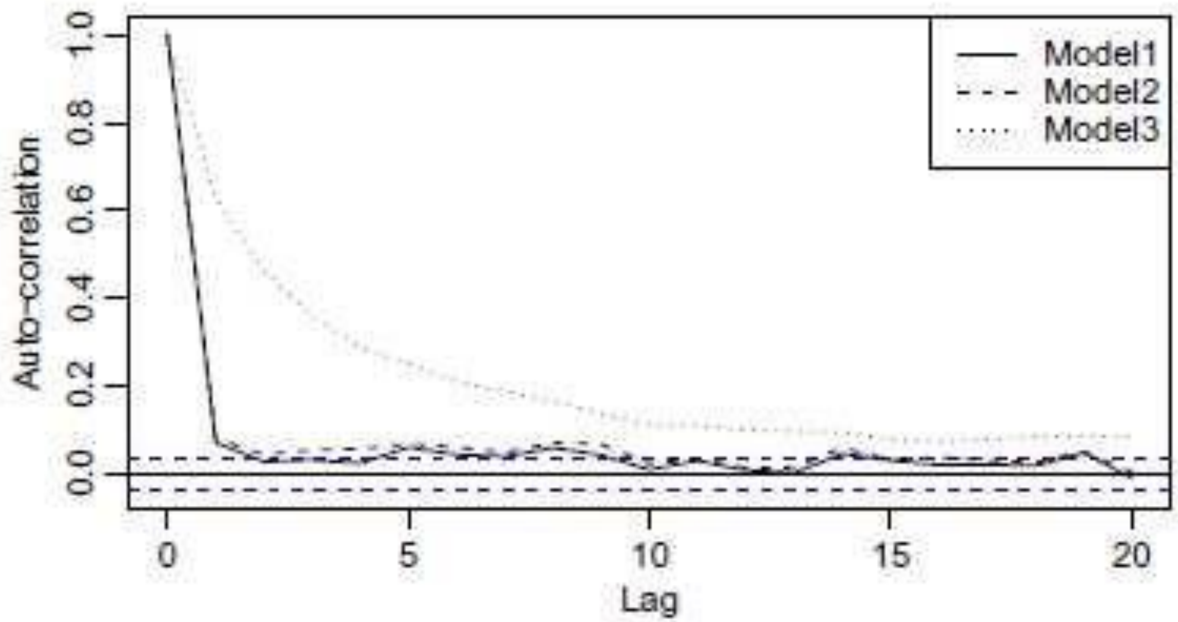
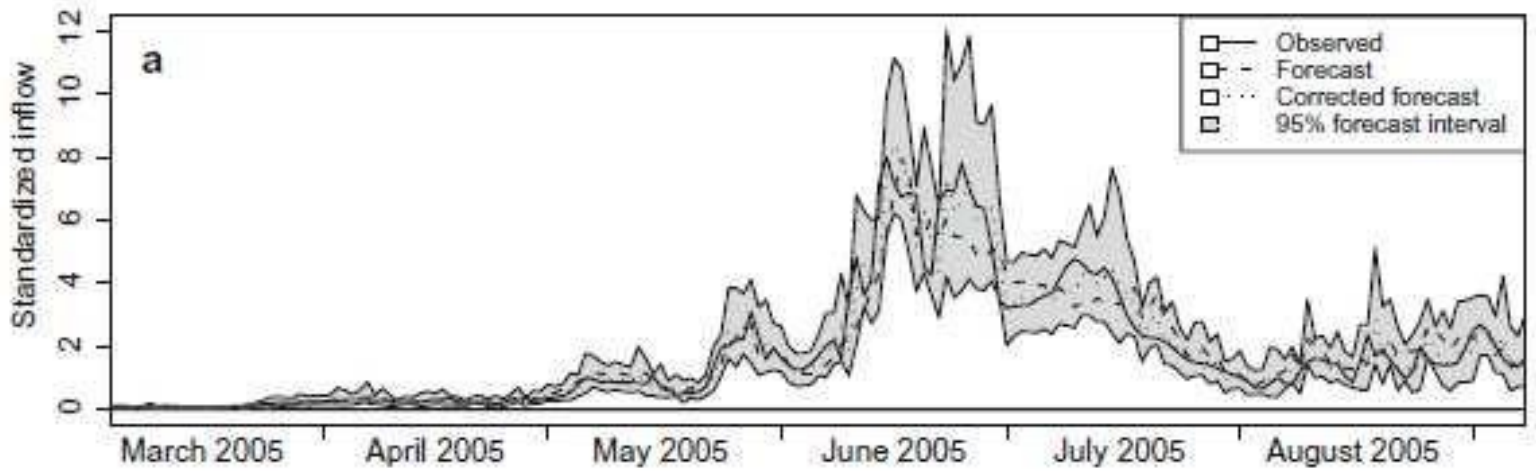


Fig. 8. Auto-correlation for p-values for Model 1 (a), Model 2 (b) and Model 3 (c).

Results


a
Observed
Forecast
Corrected forecast
95% forecast interval
Standardized inflow
0
2
4
6
8
10
12
March 2005
April 2005
May 2005
June 2005
July 2005
August 2005