Model selection, cross validation, and performance of time series models

FISH 550 – Applied Time Series Analysis

Eric Ward, warde@uw.edu

9 May 2023

Overview of today's material

- Approaches for model selection
- Cross validation
- Quantifying forecast performance

How good are our models?

Several candidate models might be built based on

- hypotheses / mechanisms
- diagnostics / summaries of fit

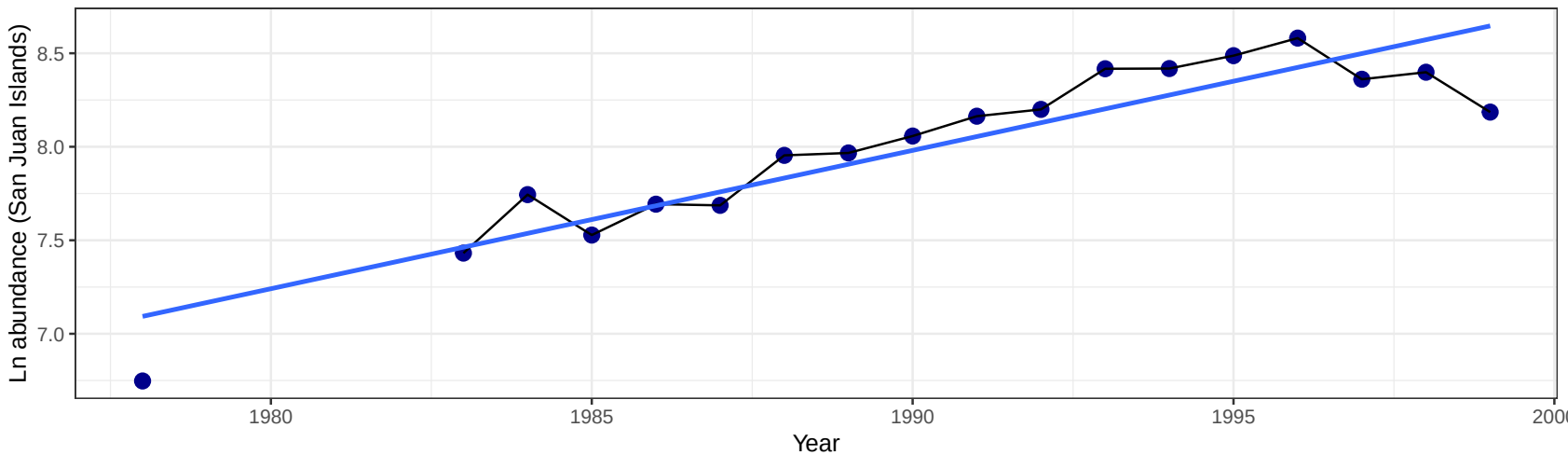
Models can be evaluated by their ability to explain data

- OR by the tradeoff in the ability to explain data, and ability to predict future data
- OR just in their predictive abilities
  - Hindcasting
  - Forecasting

How good are our models?

We can illustrate with an example to the `harborSealWA` dataset in MARSS

$$y_t = b_0 + b_1 * t + e_t$$



How good are our models?

```
##
## Call:
## lm(formula = SJI ~ Year, data = harborSealWA)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.46099 -0.08022  0.06576  0.13286  0.21464
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.392e+02  1.601e+01  -8.697 1.85e-07 ***
## Year         7.397e-02  8.043e-03   9.197 8.69e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
##
## Residual standard error: 0.1916 on 16 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.8409, Adjusted R-squared:  0.831
```

How good are our models?

Our regression model had a pretty good SS

$$SS = \sum_{i=1}^{n} (y_i - E[y_i])^2$$

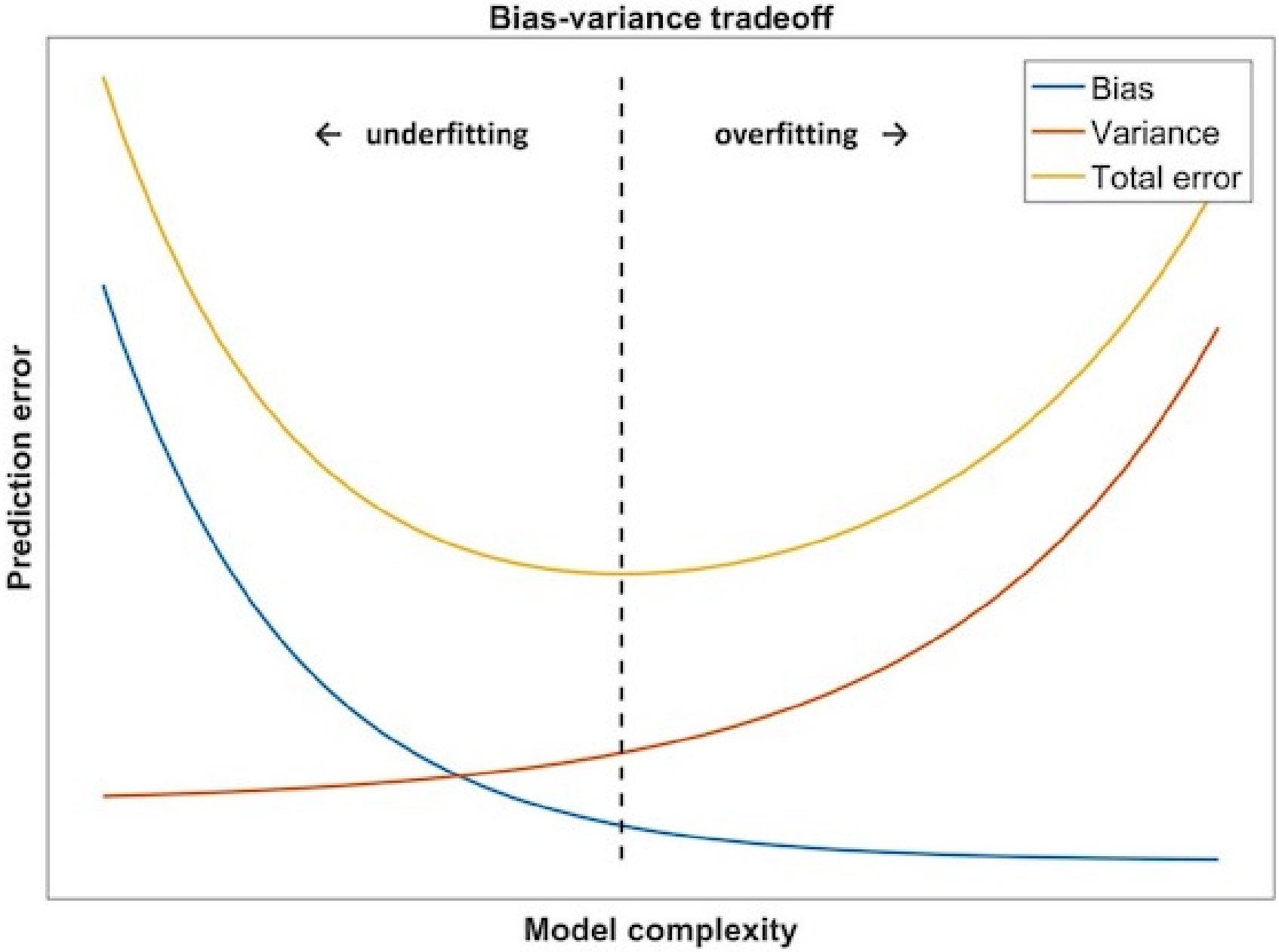
- But SS is problematic
  - as we consider more complex models, they'll inevitably reduce SS
  - there's no cost or penalty for having too many parameters

Model selection

Lots of metrics have been developed to overcome this issue and penalize complex models

- **Occam's razor**: "the law of briefness"
- **Principle of parsimony**: choose the simplest possible model that explains the data pretty well
  - choose a model that minimizes bias *and* variance

Model selection



https:

Model selection: AIC

Akaike's Information Criterion (**AIC**, Akaike 1973)

- Attempts to balance the goodness of fit of the model against the number of parameters
- Based on deviance = minus twice negative log likelihood

Deviance =

$$-2 \cdot ln\left(L(\underline{\theta}|\underline{y})\right)$$

- Deviance is a measure of model fit to data
  - lower values are better
  - Maximizing likelihood is equivalent to minimizing negative likelihood

Model selection: AIC

- Why the large focus on AIC?
- Heavily used in ecology (Burnham and Anderson 2002)[https://www.springer.com/gp/book/9780387953649]
- Also the default in many stepwise model selection procedures in R
- forecast, glmulti, bestglm, AICcmodavg, MuMIn

Model selection: AIC

Many base functions in R support the extraction of AIC

```
y = cumsum(rnorm(20))
AIC(lm(y~1))
AIC(glm(y~1))
AIC(mgcv::gam(y~1))
AIC(glmmTMB::glmmTMB(y~1))
AIC(lme4::lmer(y~1))
AIC(stats::arima(y))
AIC(forecast::Arima(y))
AIC(MARSS:MARSS(y))
```

Model selection: AIC

Many *IC approaches to model selection also rely on deviance. Where they differ is how they structure the penalty term.

For AIC, the penalty is 2 * number of parameters (k),

$$AIC = -2 \cdot ln\left(L(\underline{\theta}|\underline{y})\right) + 2k$$

- This is not affected by sample size, n

Model selection: AIC

Small sample AIC

$$AICc = AIC + \frac{2k(k+1)}{n-k-1}$$

- What happens to this term as n increases?

Model selection: AIC

AIC aims to find the best model to predict data generated from the same process that generated your observations

Downside: AIC has a tendency to overpenalize, especially for more complex models

- Equivalent to significance test w/ $\alpha = 0.16$

Alternative: Schwarz/Bayesian Information Criterion (SIC/BIC)

- Not Bayesian!
- Relies on Laplace approximation to posterior
- α becomes a function of sample size

Model selection: AIC

BIC is measure of explanatory power (rather than balancing explanation / prediction)

$$BIC = -2 \cdot ln\left(L(\underline{\theta}|\underline{y})\right) + k \cdot ln(n)$$

- Tendency of BIC to underpenalize

Model selection: AIC

Philosophical differences between AIC / BIC

- AIC / AICc tries to choose a model that approximates reality
  - does not assume that reality exists in your set of candidate models
  - One
- BIC assumes that one of your models is truth
  - This model will tend to be favored more as sample size increases

AIC and BIC for time series forecasting

- Smallest AIC similar to minimizing one-step ahead forecasts using MSE Rob Hyndman's blog
- AIC approximates LOOCV Stone (1977)
- BIC approximates k-fold cross validation Shao (1997)

Bayesian model selection

The big difference between the Bayesian and maximum likelihood approaches are that

- ML methods are maximizing the likelihood over the parameter space
- Bayesian methods are integrating over the parameter space, asking 'what values are best, on average?'

Many of the ML methods discussed were designed for models with only fixed effects.

- What about correlated parameters, nested or hierarchical models?

Bayesian model selection

Again, lots of options that have evolved quickly over the last several decades

- Bayes factors (approximated by BIC)
  - can be very difficult to calculate for complex models
- Deviance Information Criterion (DIC)
  - Spiegelhalter et al. (2002)
  - DIC is easy to get out of some programs (JAGS)
  - DIC is also attempting to balance bias and variance
- Widely applicable information criterion (WAIC)
  - Watanabe (2010)
- Leave One Out Information Criterion (LOOIC)
  - Vehtari et al. 2017, Vehtari et al. 2019

Cross validation

Recent focus in ecology & fisheries on prediction

Dietze et al. 2017

Maris et al. 2017

Pennekamp et al. 2017

Pennekamp et al. 2018

Szuwalkski & Thorson 2017

Anderson et al. 2017

Resampling techniques

Jackknife

- Hold out each data point, recomputing some statistic (and then average across 1:n)

Bootstrap

- Similar to jackknife, but with resampling

Cross-validation (k-fold)

- Divide dataset into k-partitions
- How well do (k-1) partitions predict kth set of points?
- Relationship between LOOCV and AIC / BIC

Data split: test/training sets (e.g. holdout last 5 data pts)

Resampling techniques: bootstrap

Bootstrap or jackknife approaches are useful

- generally used in the context of time series models to generate new or pseudo-datasets
- posterior predictive checks in Bayesian models generate new data from posterior draws
- state space models: use estimated deviations / errors to simulate, estimate CIs

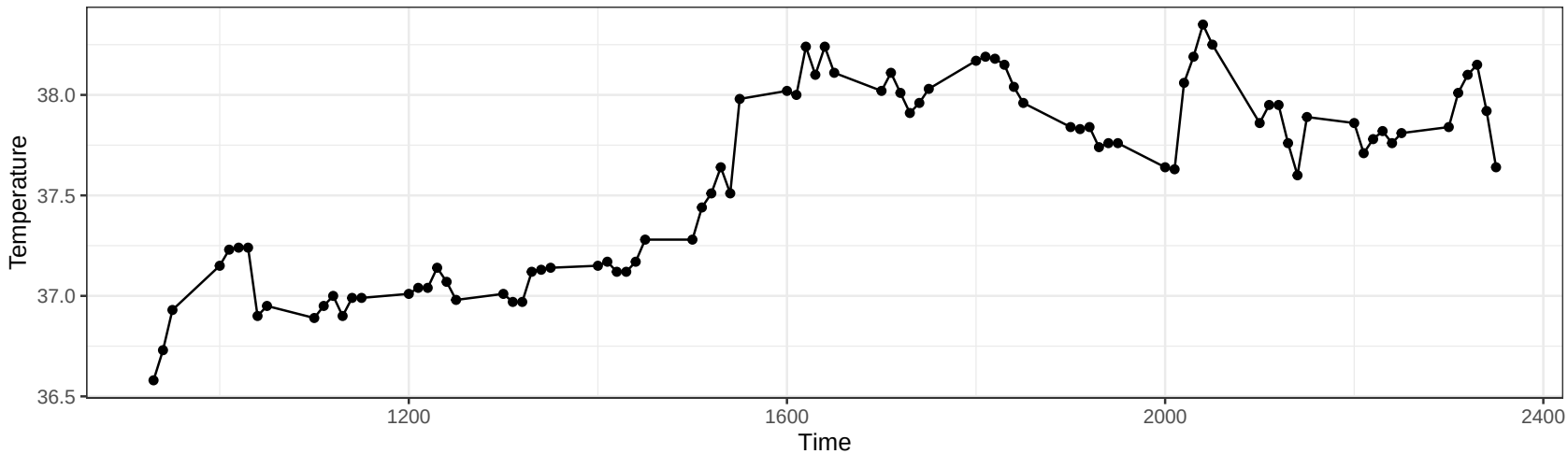
Examples

```
MARSS::MARSSboot()
MARSS::MARSSinnovationsboot()
forecast::bld.mbb.bootstrap()
forecast::forecast(..., bootstrap=TRUE)
```

Resampling techniques: K-fold cross validation

As an example, we'll use a time series of body temperature from the `beavers` dataset

```
data(beavers)
beaver = dplyr::filter(beaver2, time>200)
```



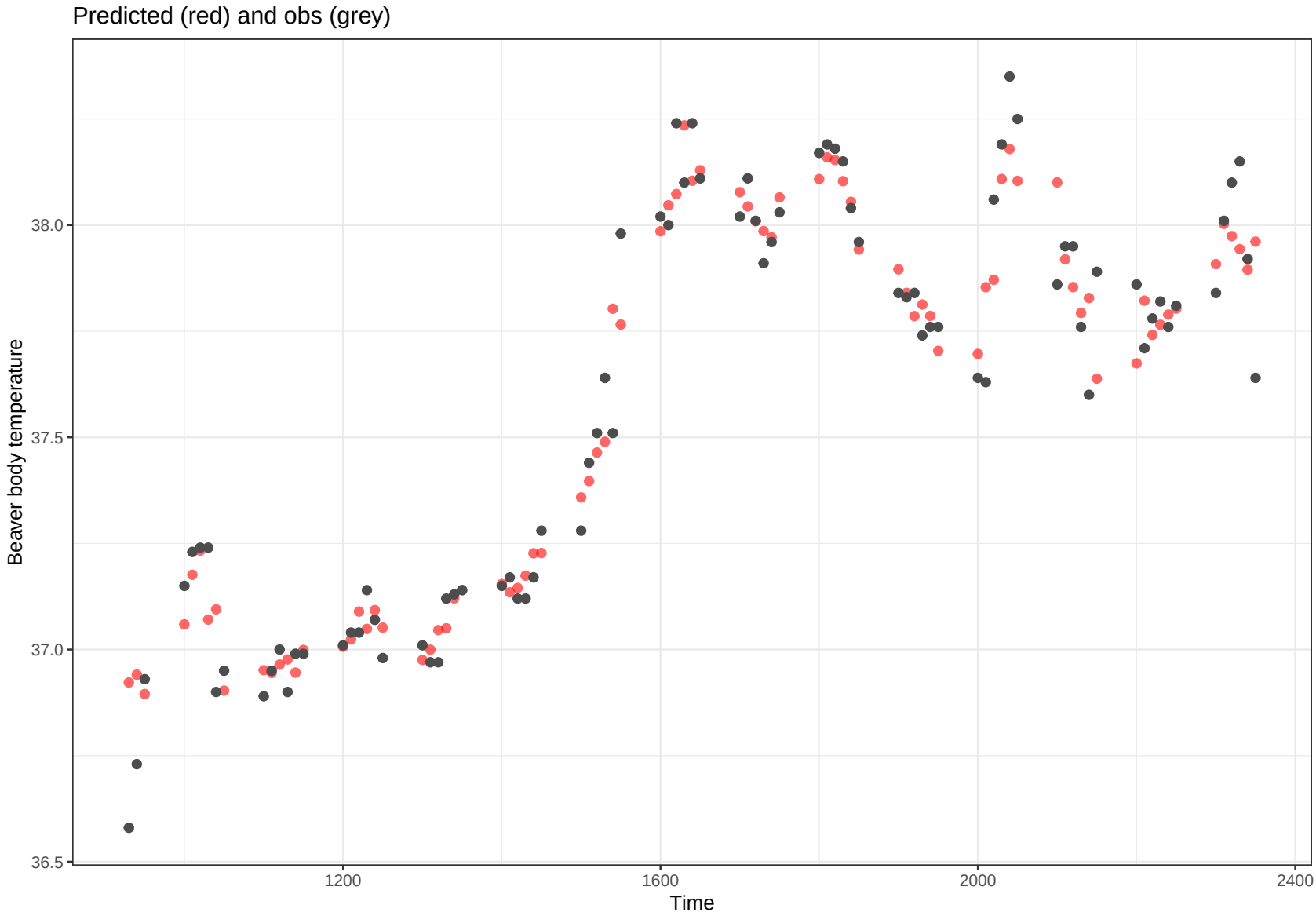
Resampling techniques: K-fold cross validation

- Choose model (e.g. State space model w/MARSS)
- Partition data
- Fit & prediction

Resampling techniques: K-fold cross validation

```
set.seed(123)
K = 5
beaver$part = sample(1:K, size=nrow(beaver), replace=T)
beaver$pred = 0
beaver$pred_se = 0
for(k in 1:K) {
  y = beaver$temp
  y[which(beaver$part==k)] = NA
    mod = MARSS(y, model=list("B"="unequal"))
    beaver$pred[beaver$part==k] =
      mod$states[1,which(beaver$part==k)]
    beaver$pred_se[beaver$part==k] =
      mod$states.se[1,which(beaver$part==k)]
}
```

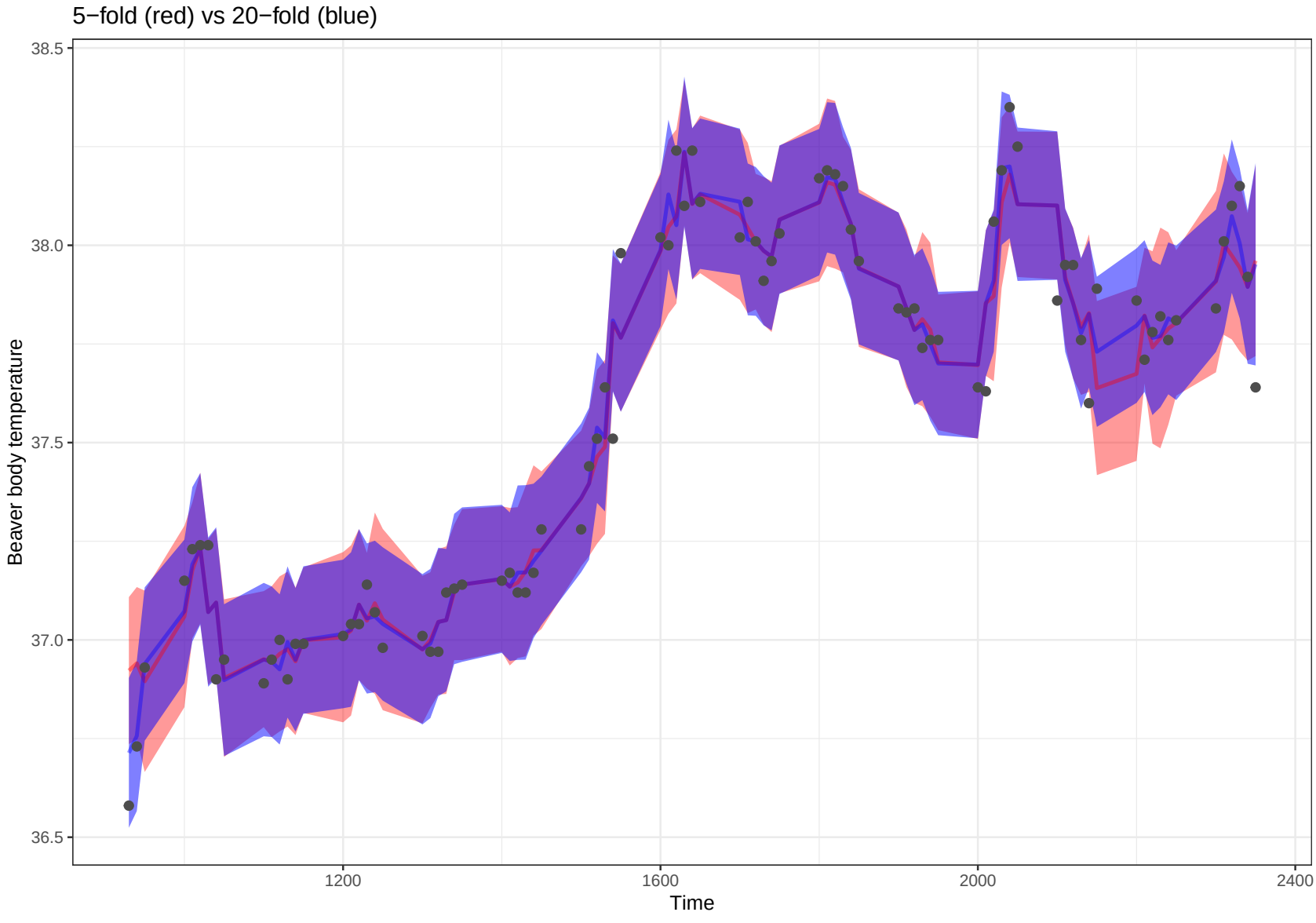
Resampling techniques: K-fold cross validation



Resampling techniques: K-fold cross validation

- How large should K be?
- Bias/variance tradeoff:
- Low K: low variance, larger bias, quicker to run. ML approaches recommend 5-10
- High K (LOOCV): low bias, high variance, computationally expensive

Resampling techniques: K-fold cross validation



Resampling techniques: repeated K-fold cross validation

- To remove effect of random sampling / partitioning, repeat K-fold cross validation and average predictions for a given data point
- caret() package in R does this for some classes of models
- Data splitting for time series

Resampling techniques: repeated K-fold cross validation

- Need to specify repeats

```
train_control = caret::trainControl(method="repeatedcv",
                    number=5, repeats=20)
```

- Again this is extendable across many widely used models

Resampling techniques

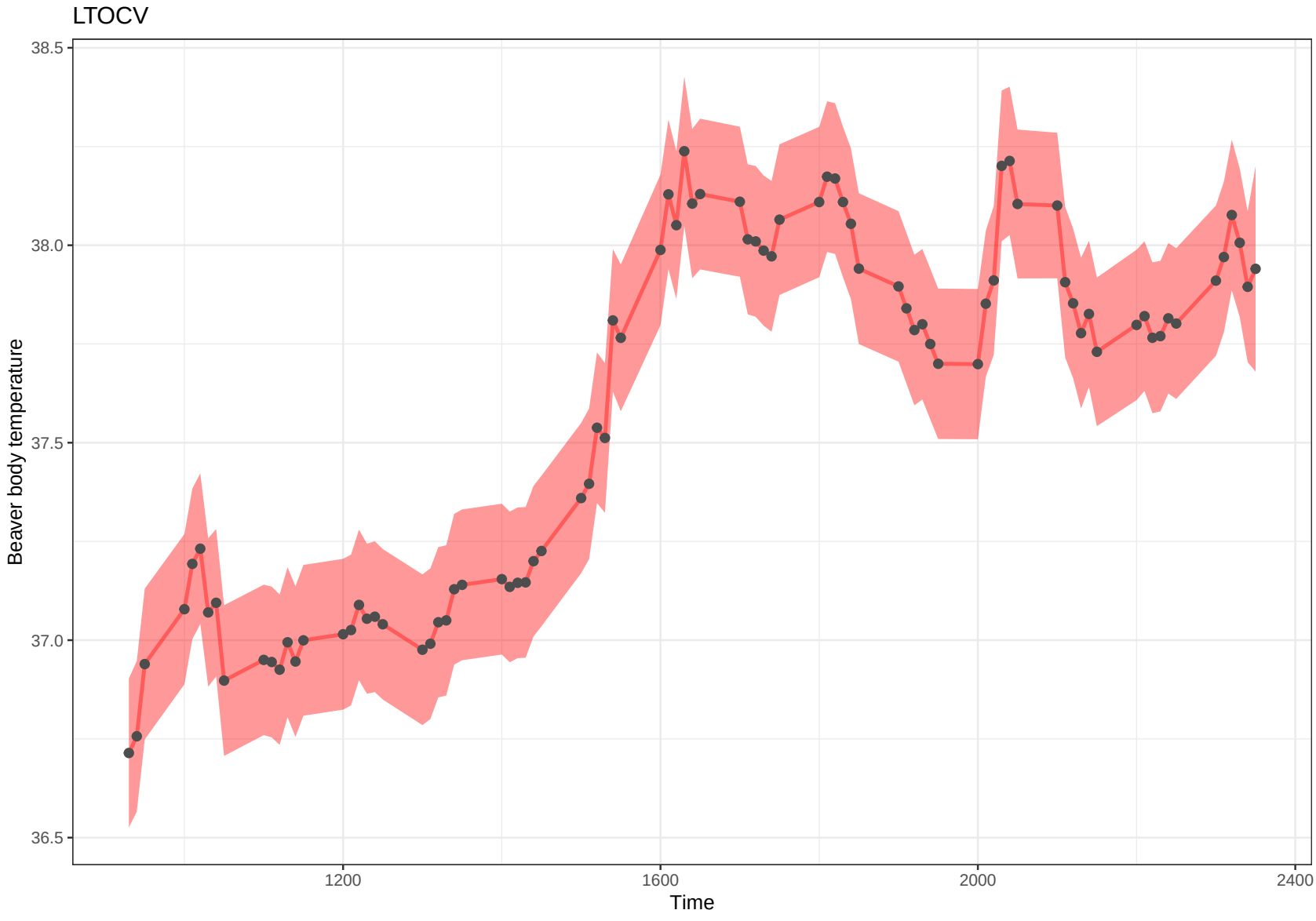
What about for time series data?

- Previous resampling was random
- No preservation of order (autocorrelation)

Resampling techniques: LTOCV

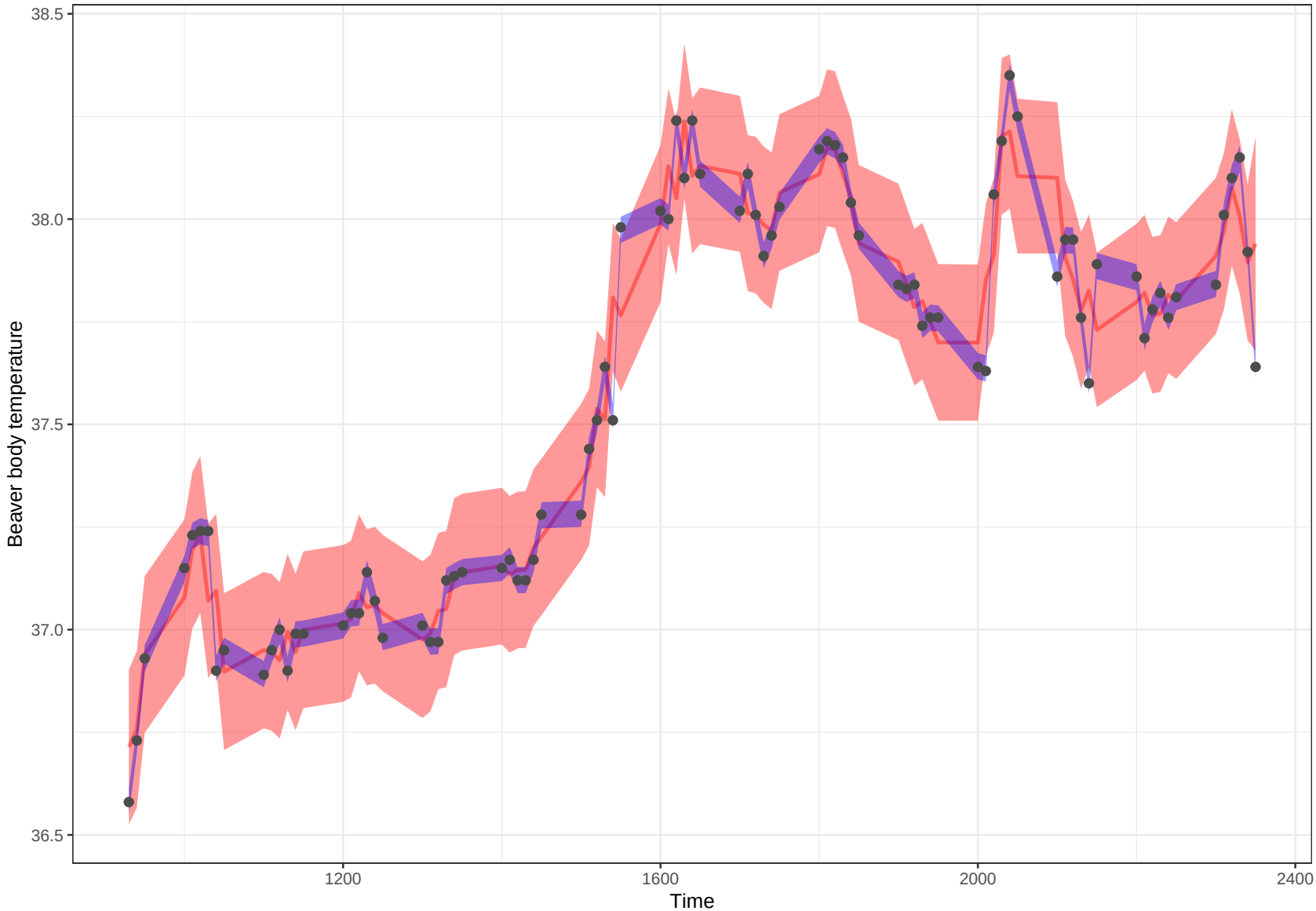
- Leave Time Out Cross Validation = leave each year out in turn
- Predict using historical and future data
- Re-analyze the beaver data using LTOCV

Resampling techniques: LTOCV



Resampling techniques: LTOCV

- Compare fit to full dataset



Resampling techniques: future (aka forward chain) CV

Leave Future Out Cross Validation: only evaluate models on future data

- Fold 1: training[1], test[2]
- Fold 2: training[1:2], test[3]
- Fold 3: training[1:3], test[4]
- Fold 4: training[1:4], test[5]

Resampling techniques: LFOCV

- Apply MARSS model to beaver dataset
- Assign partitions in order, 1:5

```
beaver$part = ceiling(5*seq(1,nrow(beaver)) / (nrow(beaver)
```

- iterate through 2:5 fitting the model and forecasting

Resampling techniques: LFOCV

Bayesian cross validation

LOOIC (Leave-one out cross validation) + preferred over alternatives

WAIC (widely applicable information criterion)

- Both available in `loo::loo()`

Additional reading: https://cran.r-project.org/web/packages/loo/vignettes/loo2-example.html

Bayesian cross validation

- Why do we need to use anything BUT `loo::loo()`?
- LOOIC is an approximation (based on importance sampling) that can be unstable for flexible (read: state space) models
- Diagnostic: Pareto-k diagnostic, 1 value per point.
- "measure of each observation's influence on the posterior of the model"
- ?diagnostics
- Stan forums or or here
- Often need to write code ourselves

Bayesian cross validation

- ELPD (Expected log posterior density)

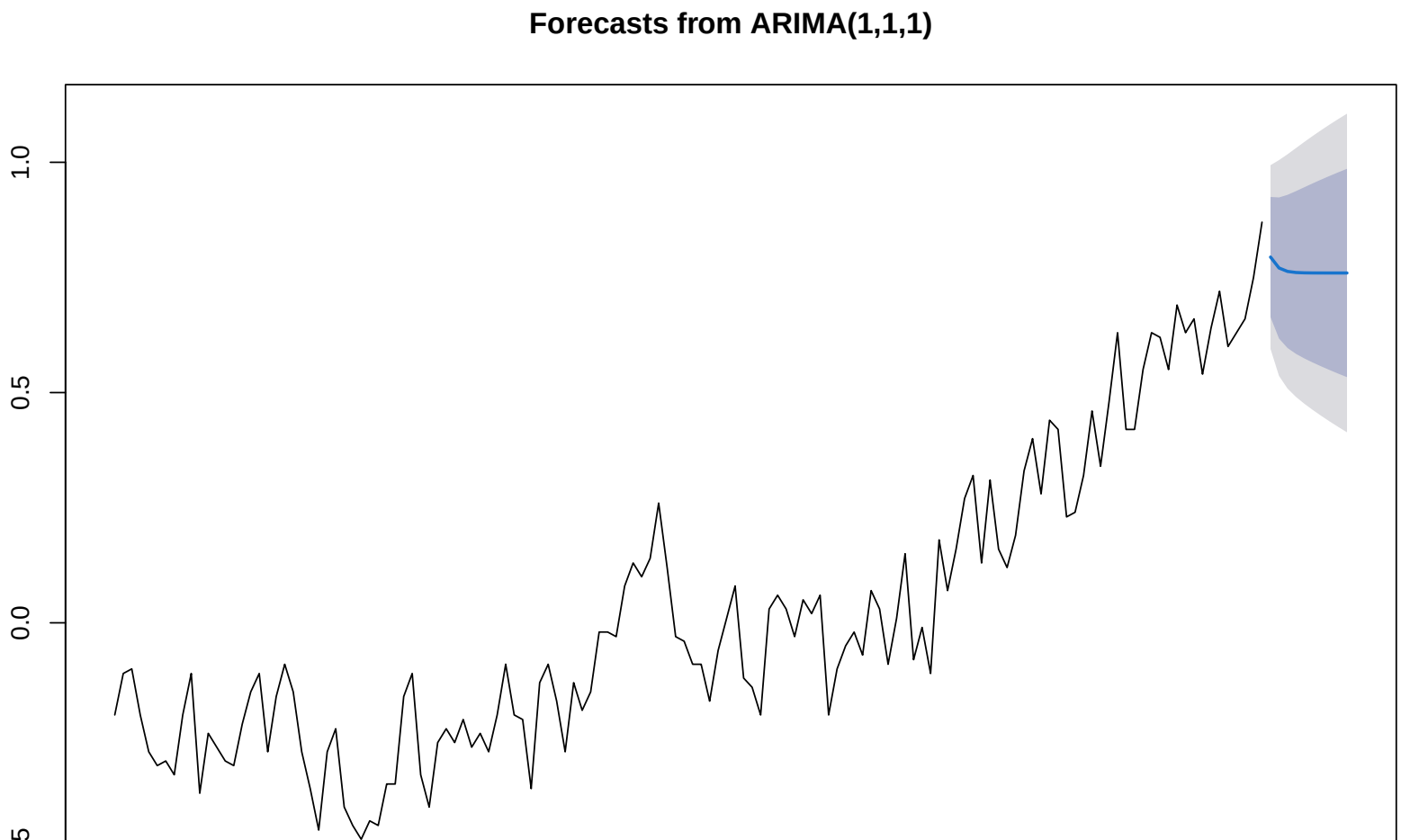
$$log[p(y^*)] = log[\int p(y^*|\theta)p(\theta)d\theta]$$

- Useful for calculating predictive accuracy for out of sample point (LTOCV, LFOCV)
- Should act similar to AIC when posterior $\sim$ MVN (more here)

Prediction and forecast evaluations

- Let's fit an ARMA(1,1) model to the global temperature data, after 1st differencing to remove trend

```
plot(f1)
```



Quantifying forecast performance

One of the most widelty used metrics is mean square error (MSE)

$$MSE = E\left[e_t^2\right] = E\left[(x_t - \hat{x}_t)^2\right]$$

- Root mean squared error (RMSE) also very common

Quantifying forecast performance

Like with model selection, the bias-variance tradeoff is important

- principle of parsimony

MSE can be rewritten as

$$MSE = Var\left(\hat{x}_t\right) + Bias(x_t, \hat{x}_t)^2$$

* Smaller MSE = lower bias + variance

Quantifying forecast performance

MSE and all forecast metrics can be calculated for

- single data points
- entire time series
- future forecasts

$$MSE = \frac{\sum_{t=1}^{n} (x_t - \hat{x}_t)^2}{n}$$

- Do you care just about predicting the final outcome of a forecast, or also the trajectory to get there?

Variants of MSE

Root mean square error, RMSE (quadratic score)

- RMSE $= \sqrt{RMSE}$
- on the same scale as the data
- also referred to as RMSD, root mean square deviation

Mean absolute error, MAE (linear score)

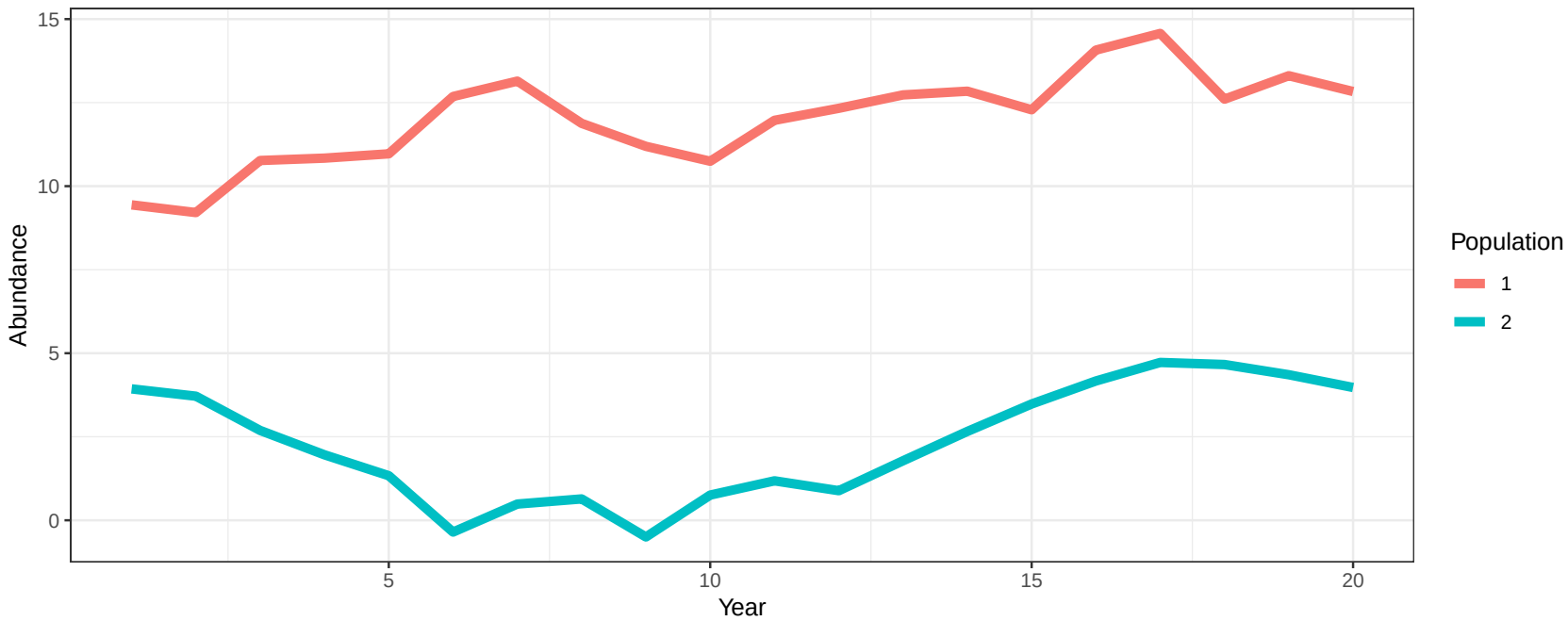
$$E\left[|x_t - \hat{x}_t|\right]$$

Median absolute error, MdAE

$$median\left[|x_t - \hat{x}_t|\right]$$

Scale independent measures of performance

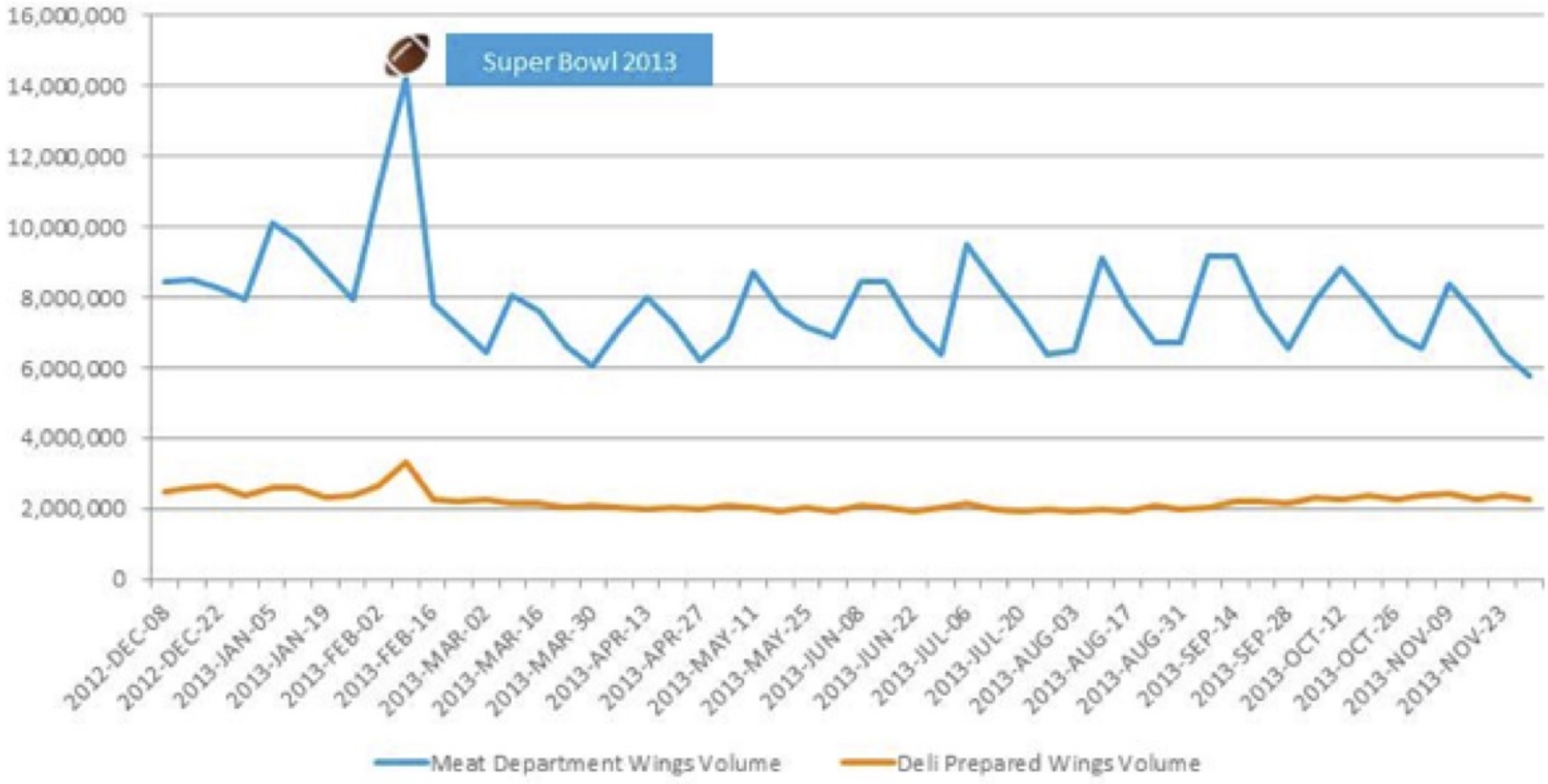
Better when applying statistics of model(s) to multiple datasets
MSE or RMSE will be driven by time series that is larger in magnitude



WING SALES SPIKED THE WEEK OF THE SUPER BOWL

Consumers also stocked up in the week leading up to the Super Bowl

Wings weekly volume, total U.S.



Meat Department Wings includes: fresh wings, fresh value-added wings and fully cooked wings

Percent Error Statistics

Percent Error:

$$p_t = \frac{e_t \cdot 100}{Y_t}$$

Mean Absolute Percent Error (MAPE):

$$MAPE = E\left[|p_t|\right]$$

Root Mean Square Percent Error (RMSPE):

$$RMSPE = \sqrt{E\left[p_t^2\right]}$$

Issues with percent error statistics

$$p_t = \frac{e_t \cdot 100}{Y_t}$$

- What happens when Y = 0?
- Distribution of percent errors tends to be highly skewed / long tails
- MAPE tends to put higher penalty on positive errors
- See Hyndman & Koehler (2006)

Scaled error statistics

Define scaled error as

$$q_t = \frac{e_t}{\frac{1}{n-1}\sum_{i=2}^{n}(Y_i - Y_{i-1})}$$

- denominator is MAE from random walk model, so performance is gauged relative to that
- this does not allow for missing data

Absolute scaled error (ASE)

$$ASE = |q_t|$$

Mean absolute scaled error (MASE)

$$MASE = E\left[|q_t|\right]$$

Interpreting ASE and MASE

All values are relative to the naïve random walk model

- Values < 1 indicate better performance than RW model
- Values > 1 indicate worse performance than RW model

Implementation in R

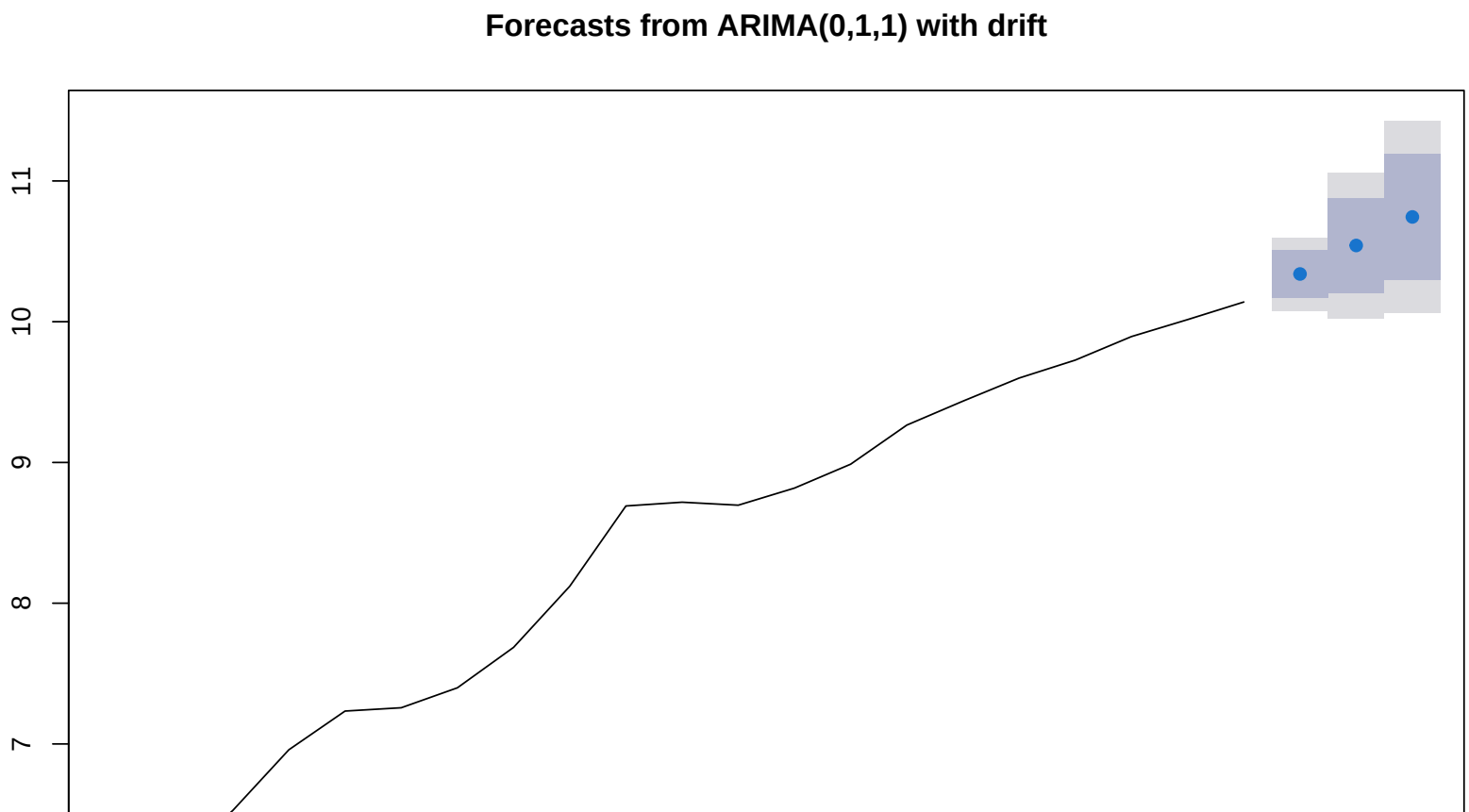
- Fit an ARIMA model to 'airmiles',holding out last 3 points

```
n = length(airmiles)
air.model = auto.arima(log(airmiles[1:(n-3)]))
```

Implementation in R

- Forecast the fitted model 3 steps ahead
- Use holdout data to evaluate accuracy

```
air.forecast = forecast(air.model, h = 3)
plot(air.forecast)
```



Implementation in R

Evaluate RMSE / MASE statistics for 3 holdouts

```
accuracy(air.forecast, log(airmiles[(n-2):n]), test = 3)
```

```
##                      ME      RMSE       MAE        MPE      MA
## Test set -0.4183656 0.4183656 0.4183656 -4.051598 4.0515
```

Evaluate RMSE / MASE statistics for only last holdout

```
accuracy(air.forecast, log(airmiles[(n-2):n]), test = 1)
```

```
##                      ME      RMSE       MAE        MPE      MA
## Test set -0.1987598 0.1987598 0.1987598 -1.960106 1.960
```

MSE vs MAPE vs MASE

Raw statistics (e.g. MSE, RMSE) shouldn't be applied for data of different scale

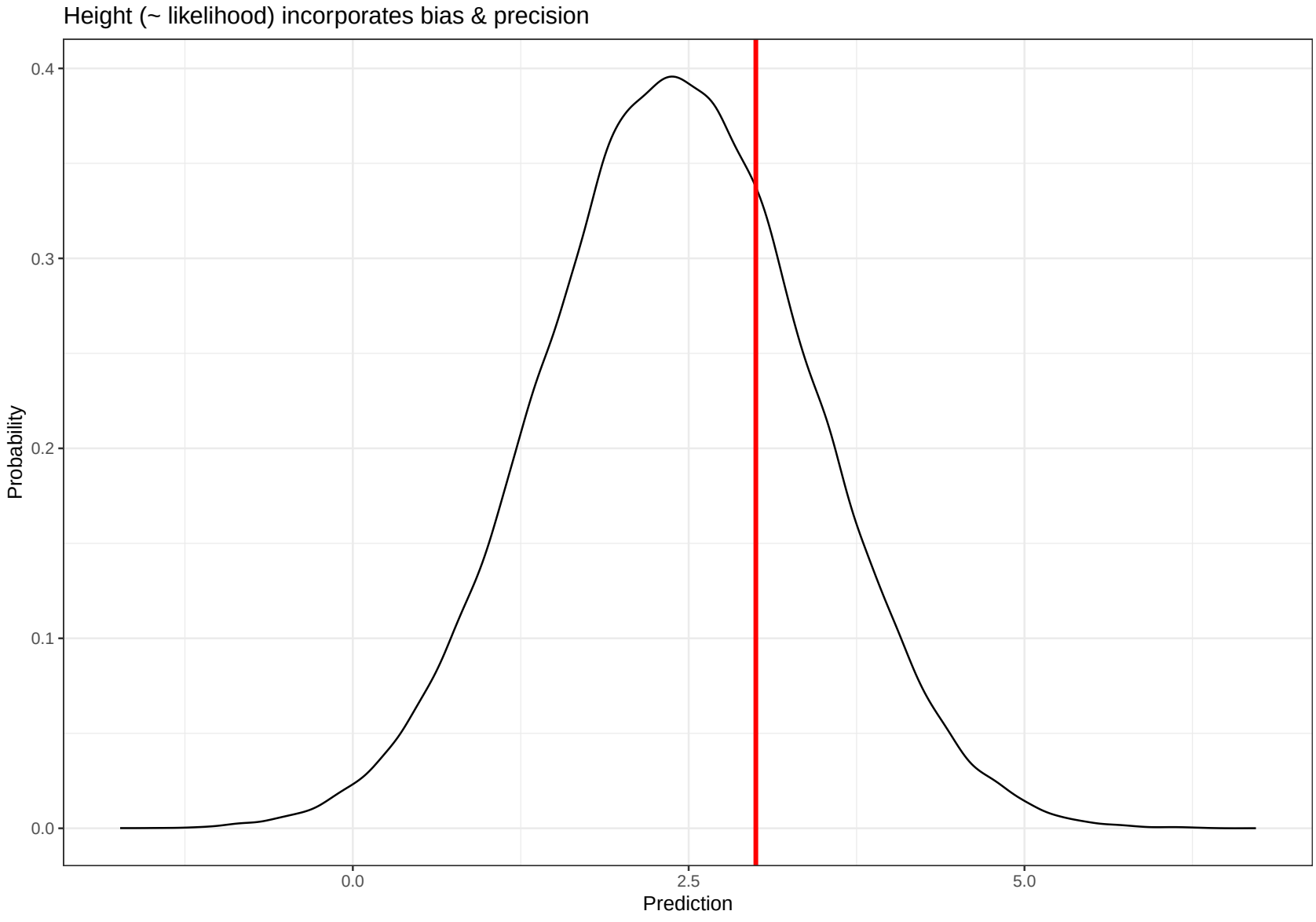
Percent error metrics (e.g. MAPE) may be skewed & undefined for real zeroes

Scaled error metrics (ASE, MASE) have been shown to be more robust meta-analyses of many datasets + Hyndman & Koehler (2006)

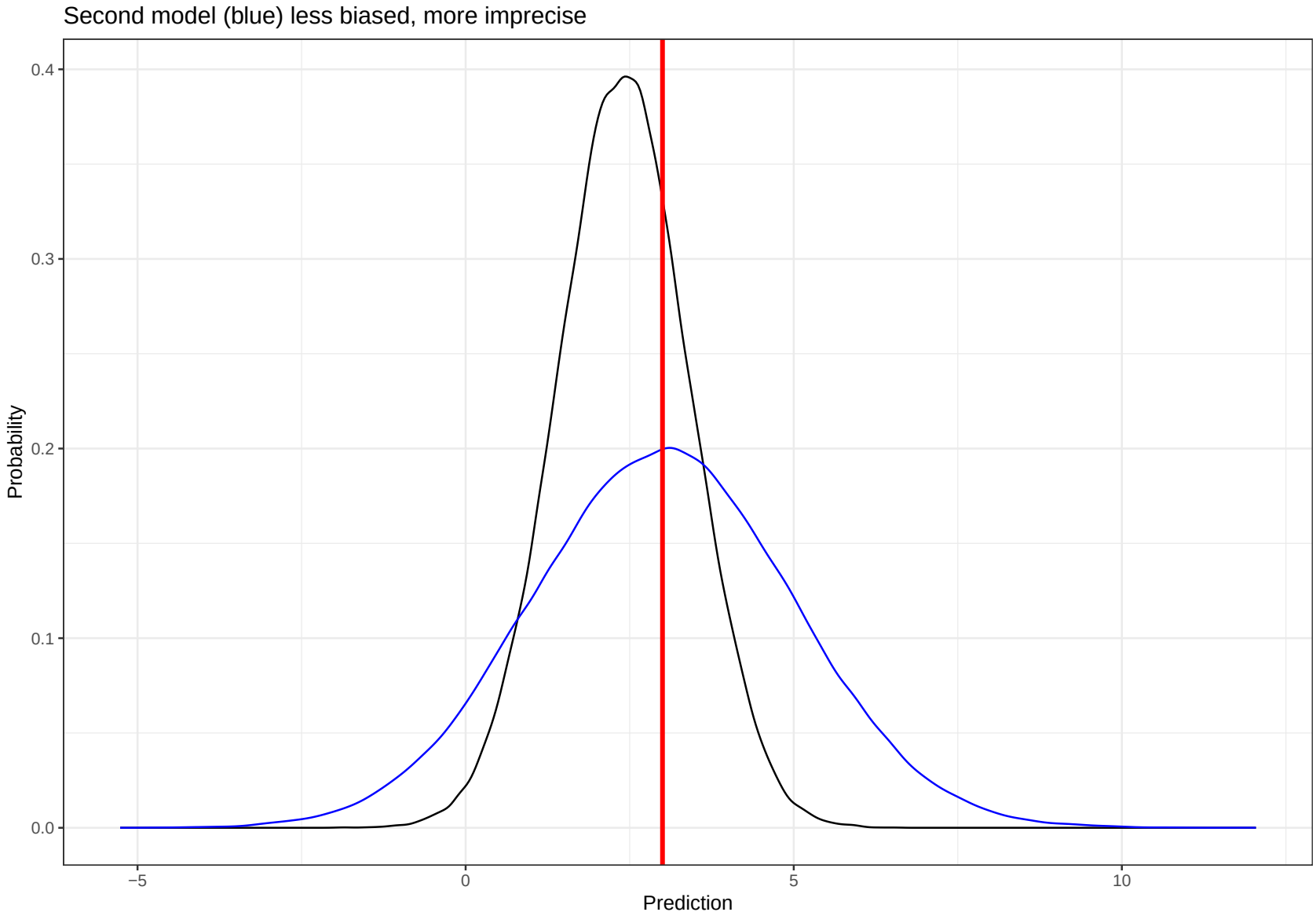
Scoring rules

- Metrics (RMSE, etc.) evaluate point estimates of predictions vs. observations
- But what if we also care about how uncertain our predictions / forecasts are?
- limited to applications of parametric methods
- Scoring rules
- Draper (2005)
- Gneting and Raftery (2012)
- R packages: scoring, scoringRules

Scoring rules



Scoring rules



Questions?