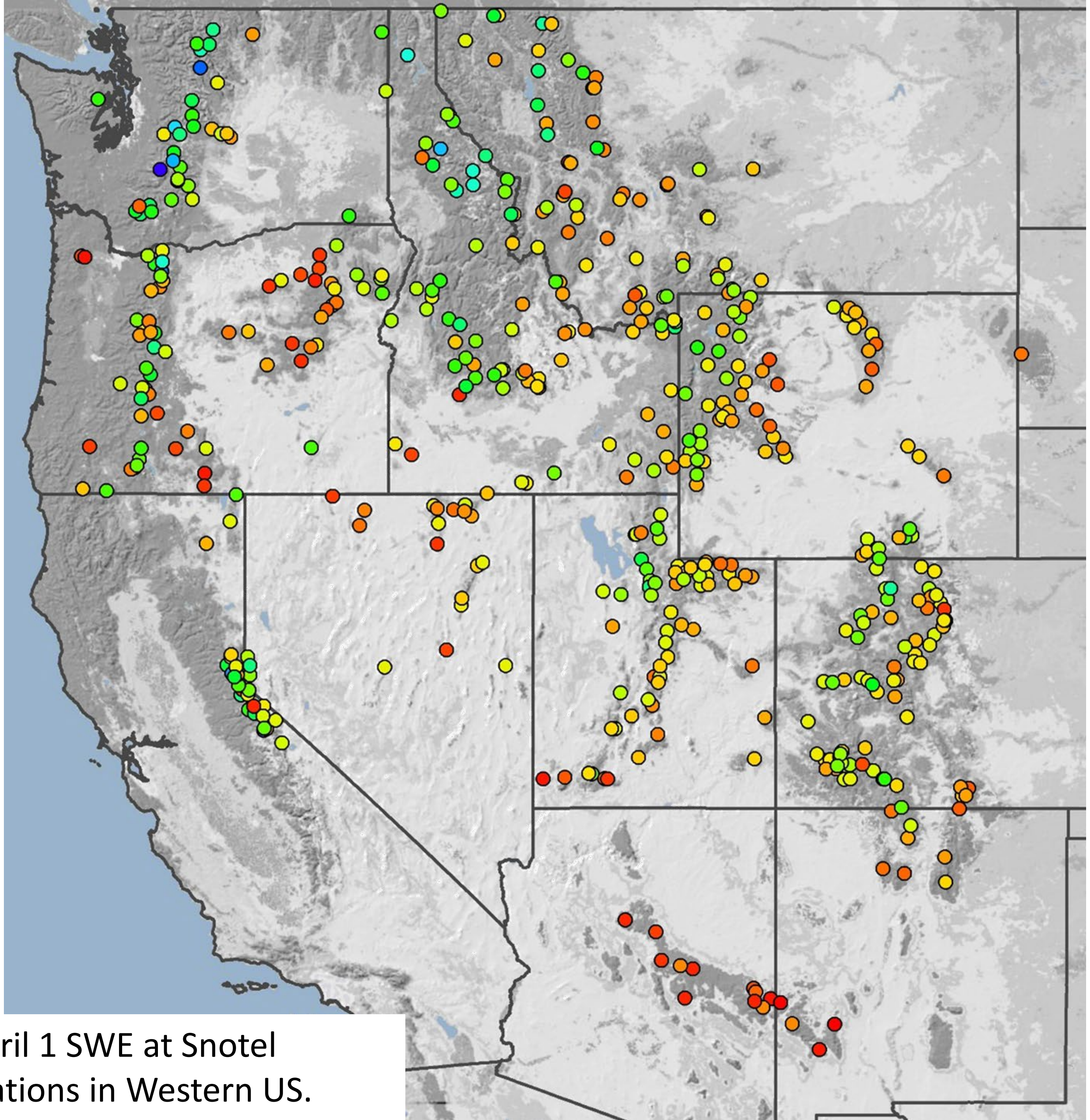


April 1 SWE at Snotel Stations in Western US.

Applying Non-Random Block Cross-Validation to Improve Reliability of Model Selection and Evaluation in Hydrology: An illustration using an algorithmic model of seasonal snowpack

Charles Luce
US Forest Service Research

A.C. Lute
University of Idaho

Surface Question: Is a model of snow based on climatological time scale information robust for transfer to other times and places?

Deeper Question: How does that transferability change with model complexity?

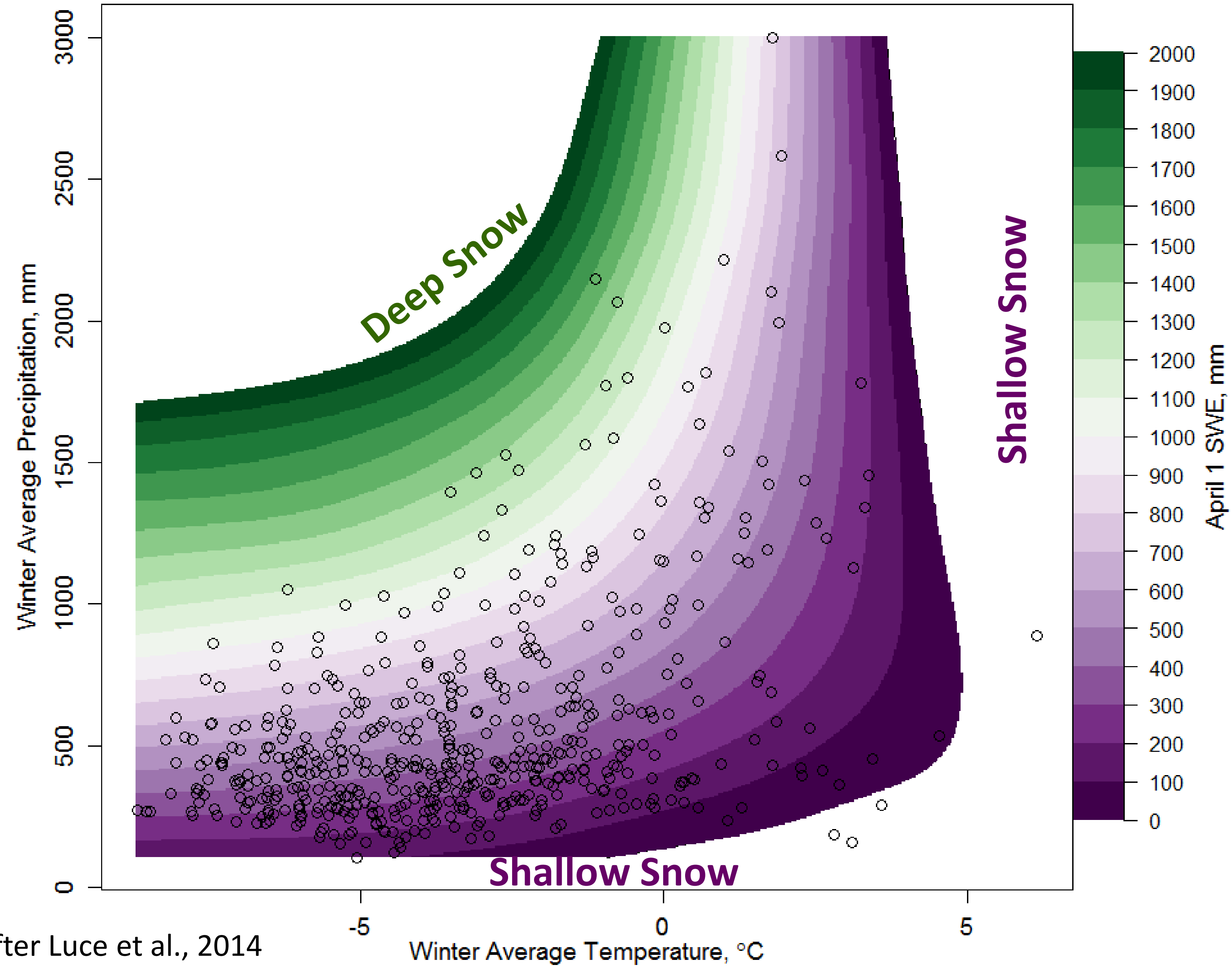
Further discussion of results in Lute and Luce, 2017

The Upshot:

Choices regarding sample and treatment of the sample in the training and validation exercise affect

1. Selection of the model structure, and
2. Evaluation of the model

OR:
Model structural uncertainty and performance depend on the data you use and how you array those data to test the alternative models.



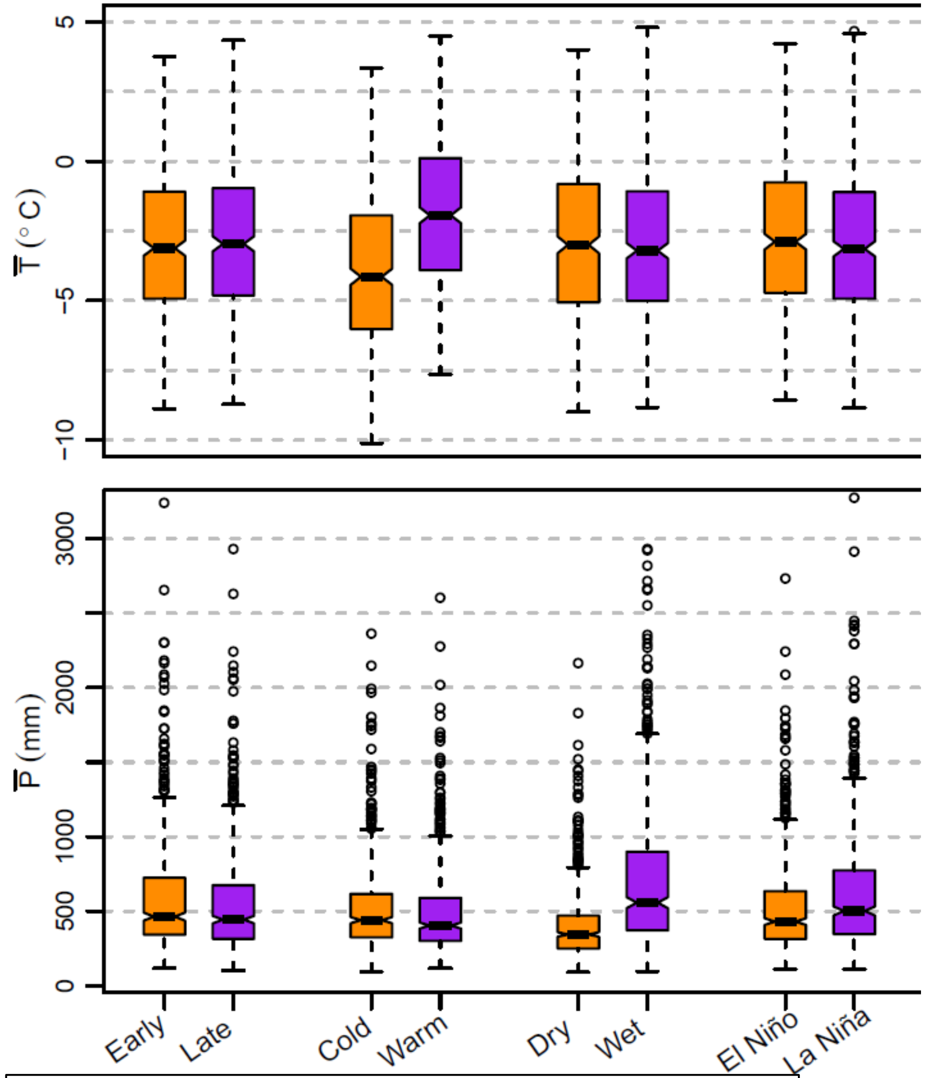
After Luce et al., 2014

The basic form of the model is

$$SWE = f(\overline{T}, P_{win})$$

Where $\overline{T}$ is mean Nov-Mar temperature and P_{win} is Nov-Mar Precipitation

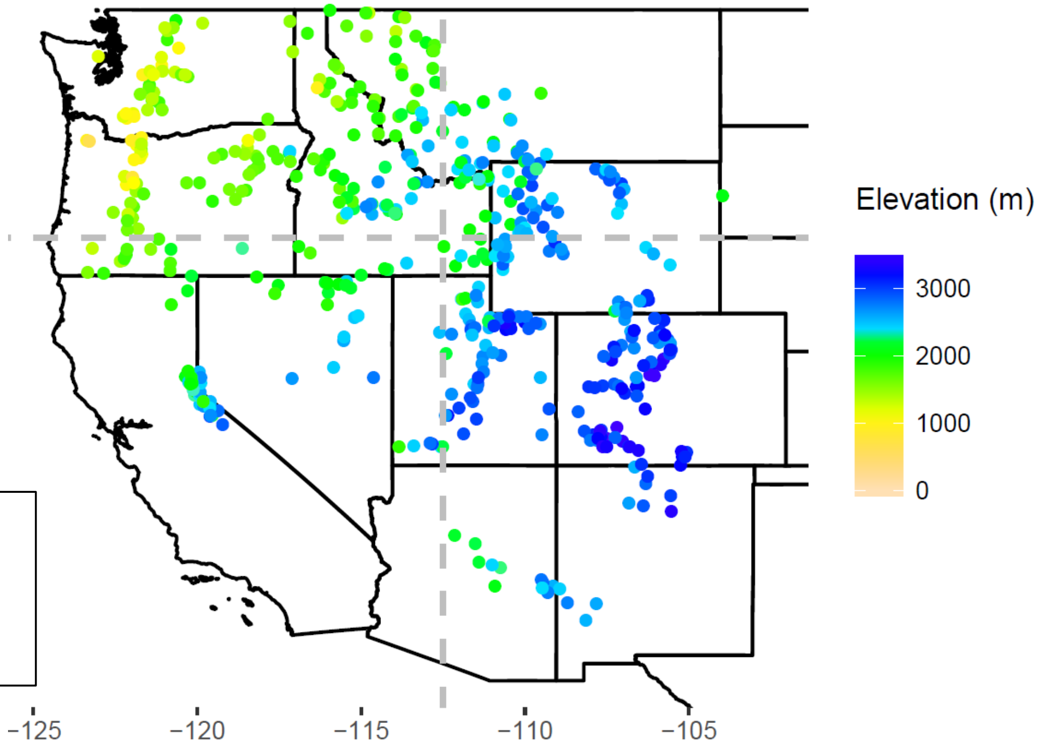
The general shape of the function is an interaction where deeper snow is generated by the combination of cold temperatures and plentiful precipitation.



Spatial and Temporal 2-Fold Block Cross Validations

Because random cross validations can underestimate errors, we applied non-random blocks to intentionally challenge the model.

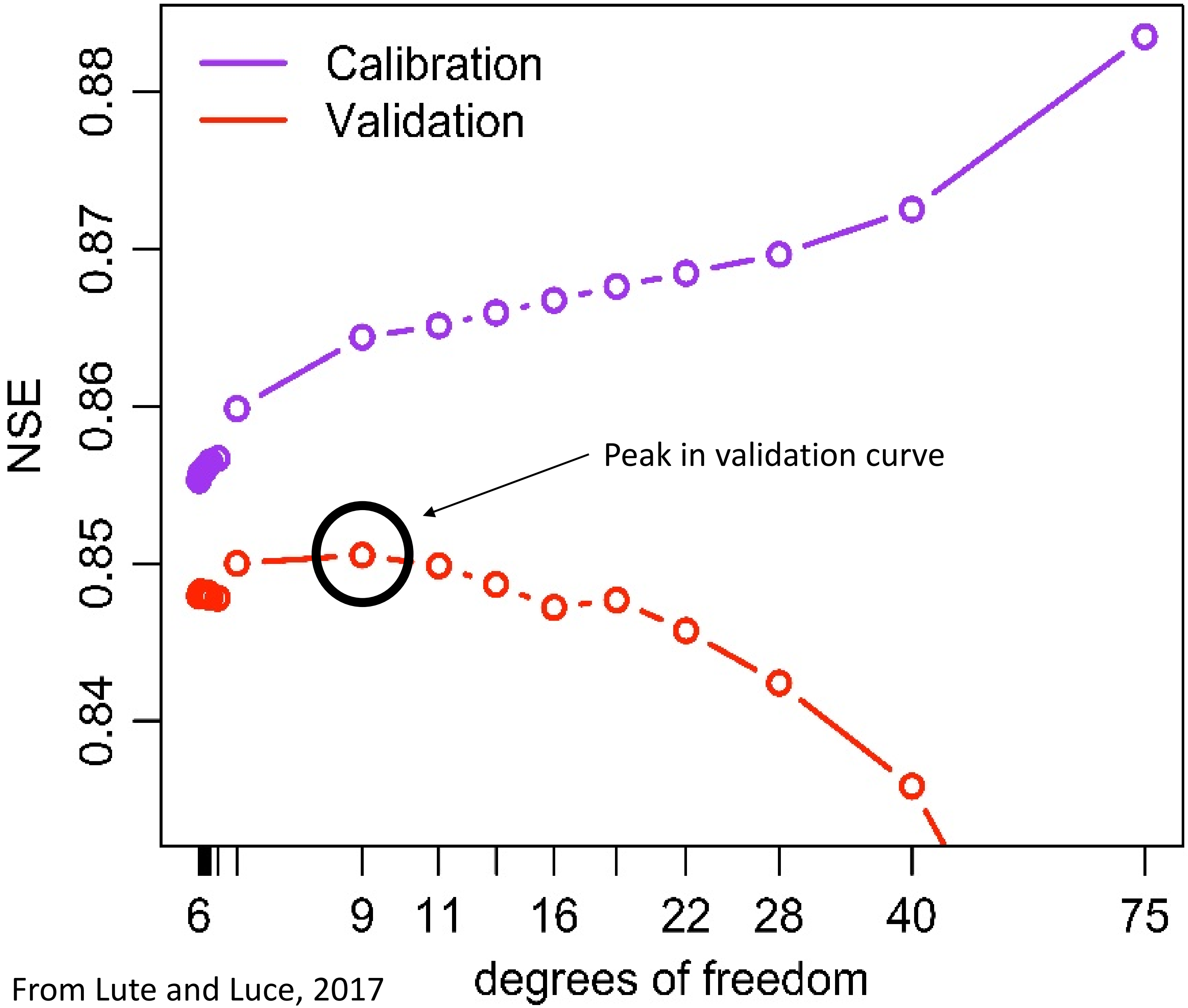
Spatial Cross-Validations: East and West and North and South Blocks
(e.g. train on east half, validate on west half)



Temporal Cross-Validations:
- First decade versus second
- Cold vs. Warm terciles
- Dry vs. Wet terciles
- ENSO tercile contrasts

Combined Space & Time Cross-Validations
e.g. Early West vs. Late East

From Lute and Luce, 2017



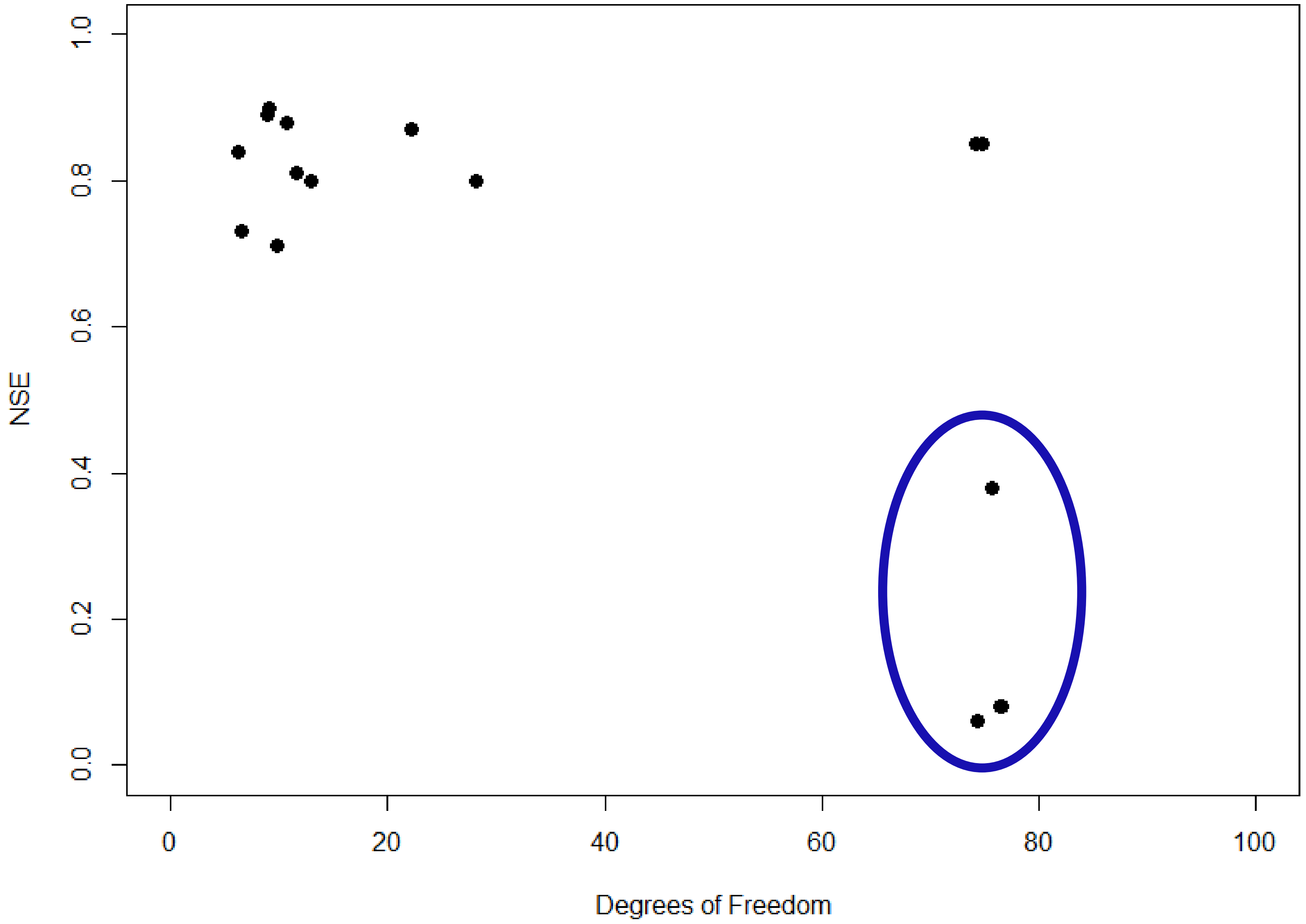
From Lute and Luce, 2017

A maximum in the validation NSE vs. complexity curve relates to bias-variance tradeoffs and shows the best performing complexity for a given training-validation combination, and the relative difference to other model complexities.

Calibration performance predictably increases with added parameters.

The difference between calibration and validation tends to be small for low numbers of parameters, when calibration is the worst.

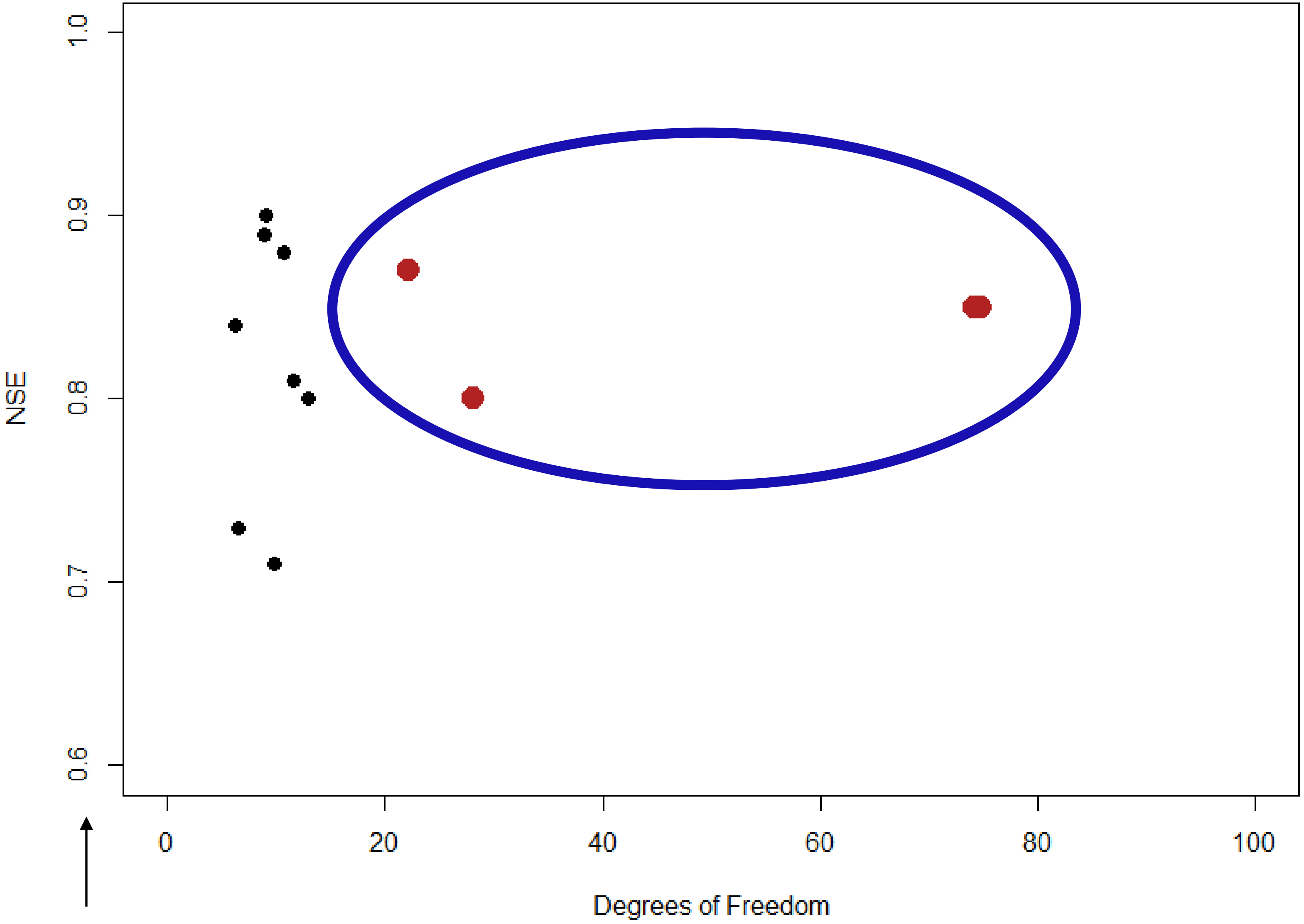
We are just interested in the validation performance peak (NSE) and sometimes neighboring points.



Plot of the “high point” in each arc of NSE versus complexity for all of the non-random block cross validations.

These four poorly performing models are a result of a restricted range in the predictors of the training set. (= too much extrapolation)

Too extreme of a differential split sample.
=> Overly pessimistic evaluation and overly complex structure



Note change
In y-axis

These high complexity models occur in temporal cross validations that have low contrast in predictor variables.

The data are not independent!
The models fit idiosyncrasies in station-level variations due to unmeasured variables.

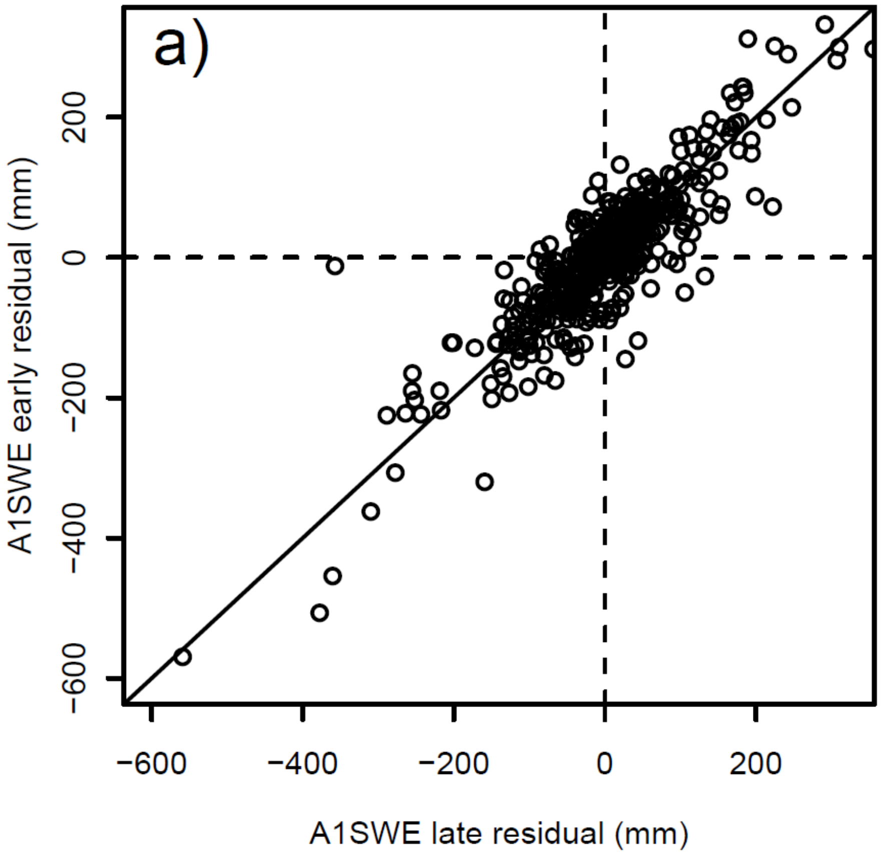
Such models attribute to T and P, things that happen because of topography, wind, or vegetation.

On Pseudoreplication:

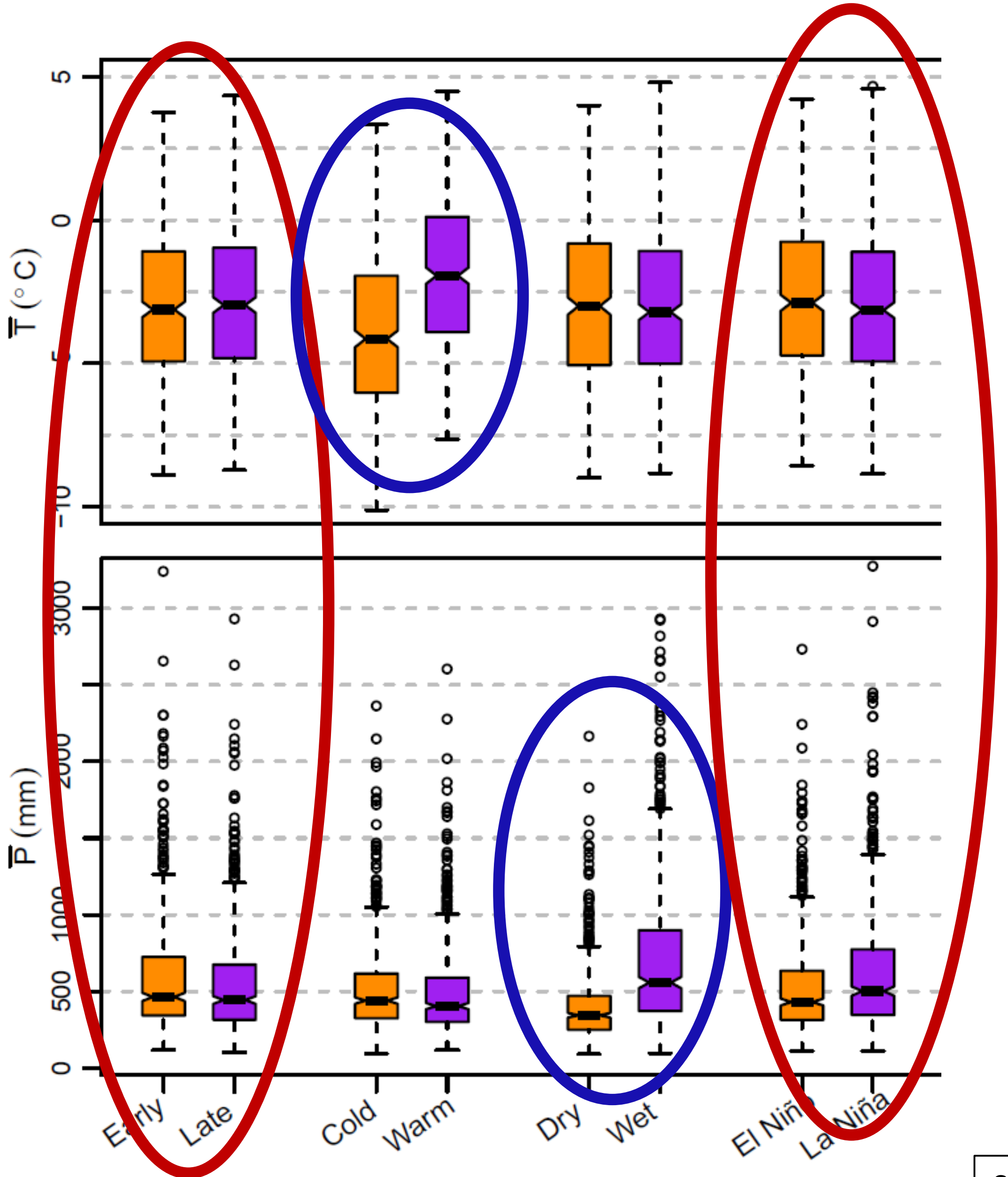
Some between-year splits are not independent unless large WX change

Red ovals show low contrast years, where high complexity models are selected.

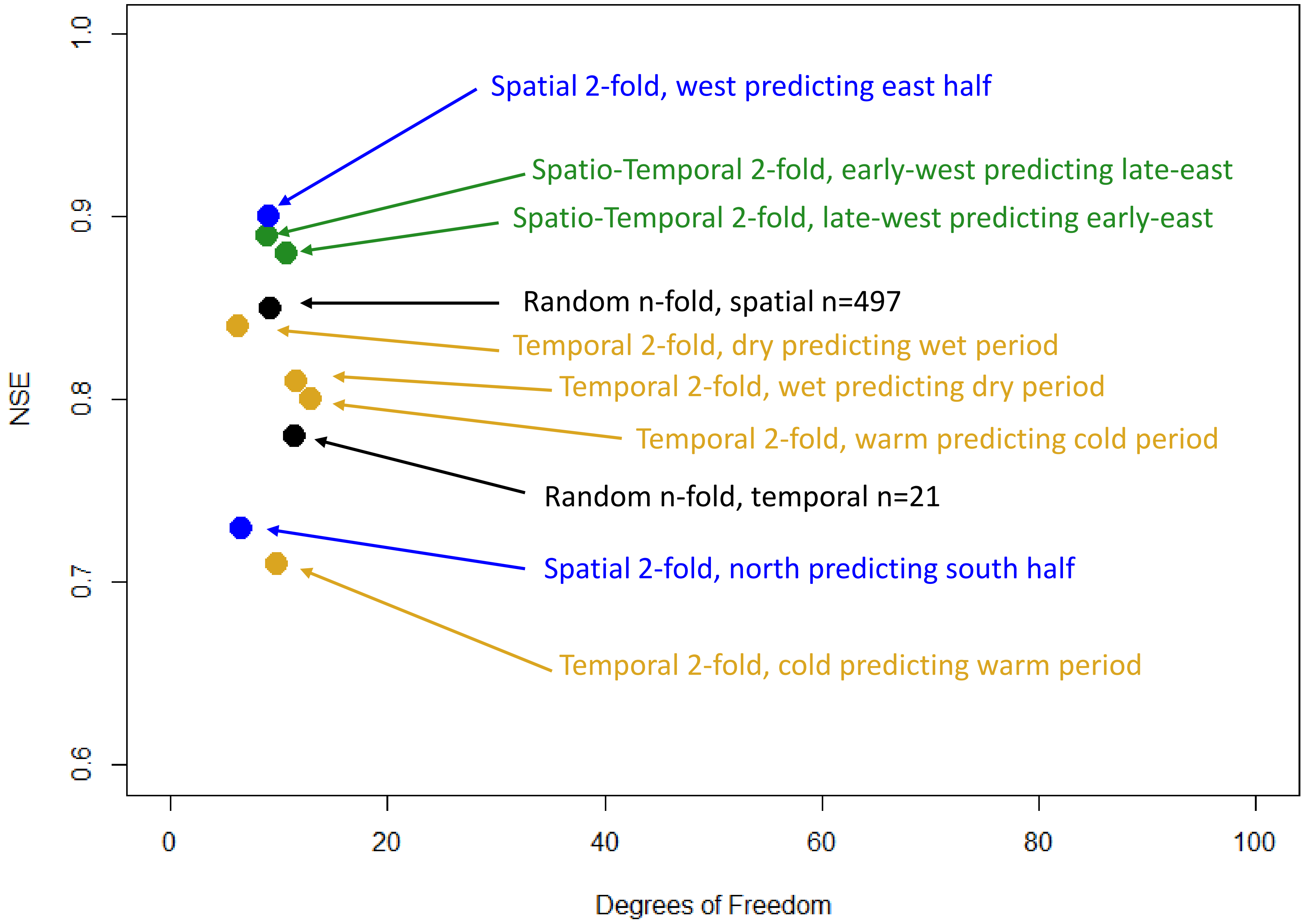
Correlated calibration errors in both data sets.



From Lute and Luce, 2017



From Lute and Luce, 2017



Substantial variation in model performance remains as a function of training and validation choices.

Fairly consistent model structure identification with low to moderate complexity.

Poorest performance elicited from cool training sets predicting warm validation sets.

The spatial random cross validation (the most obvious choice, and our initial approach) was overly optimistic.

Even Klemeš’ conceptually most strenuous test can present overoptimistic assessment.

References:

Luce, C. H., V. Lopez-Burgos, and Z. Holden (2014), Sensitivity of snowpack storage to precipitation and temperature using spatial and temporal analog models, *Water Resour. Res.*, *50*(12), 9447-9462, 10.1002/2013WR014844.

Lute, A. C., and C. H. Luce (2017), Are model transferability and complexity antithetical? Insights from validation of a variable-complexity snow model in space and time, *Water Resour. Res.*, *53*(11), 8825–8850, 10.1002/2017WR020752.