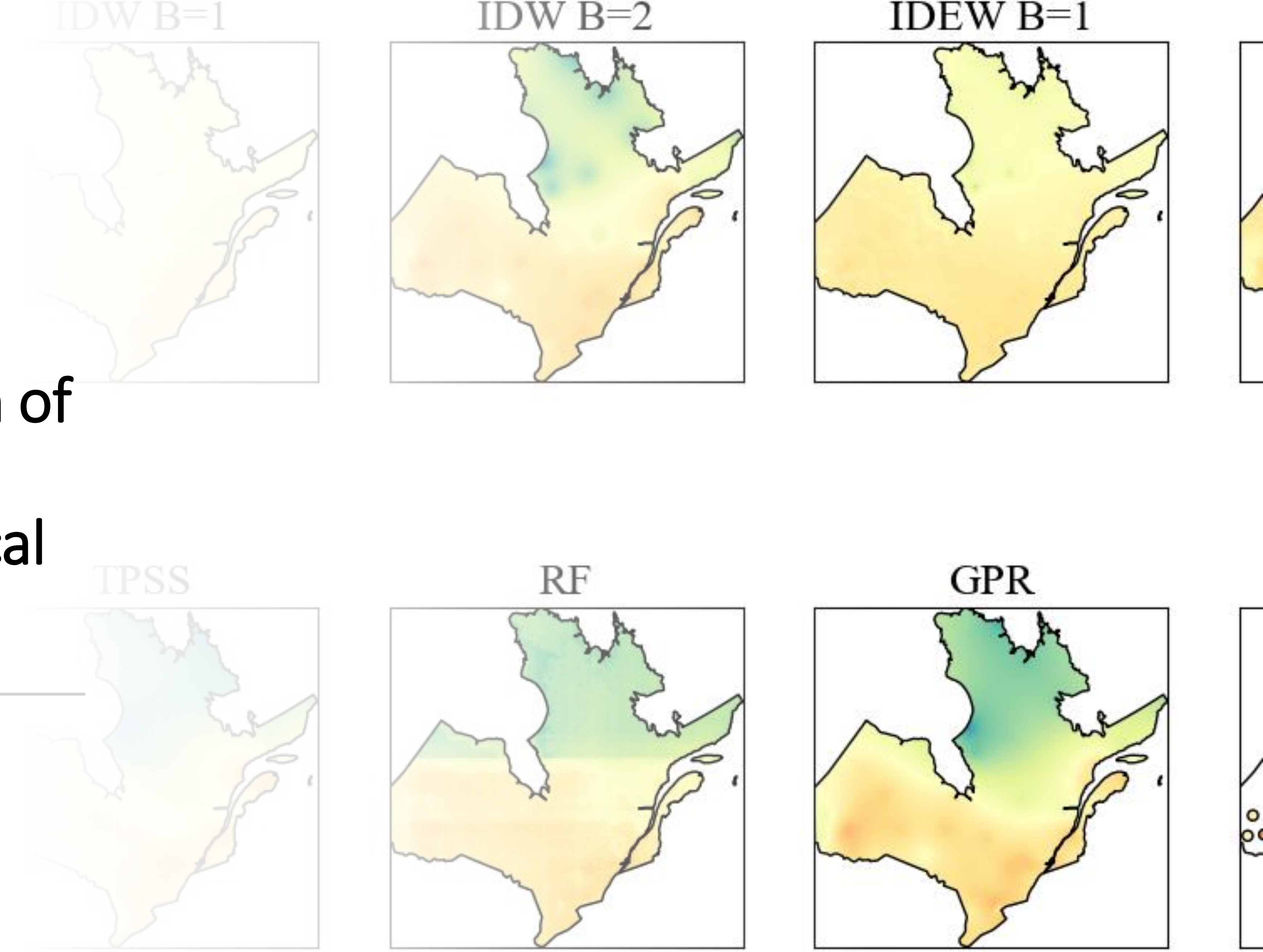
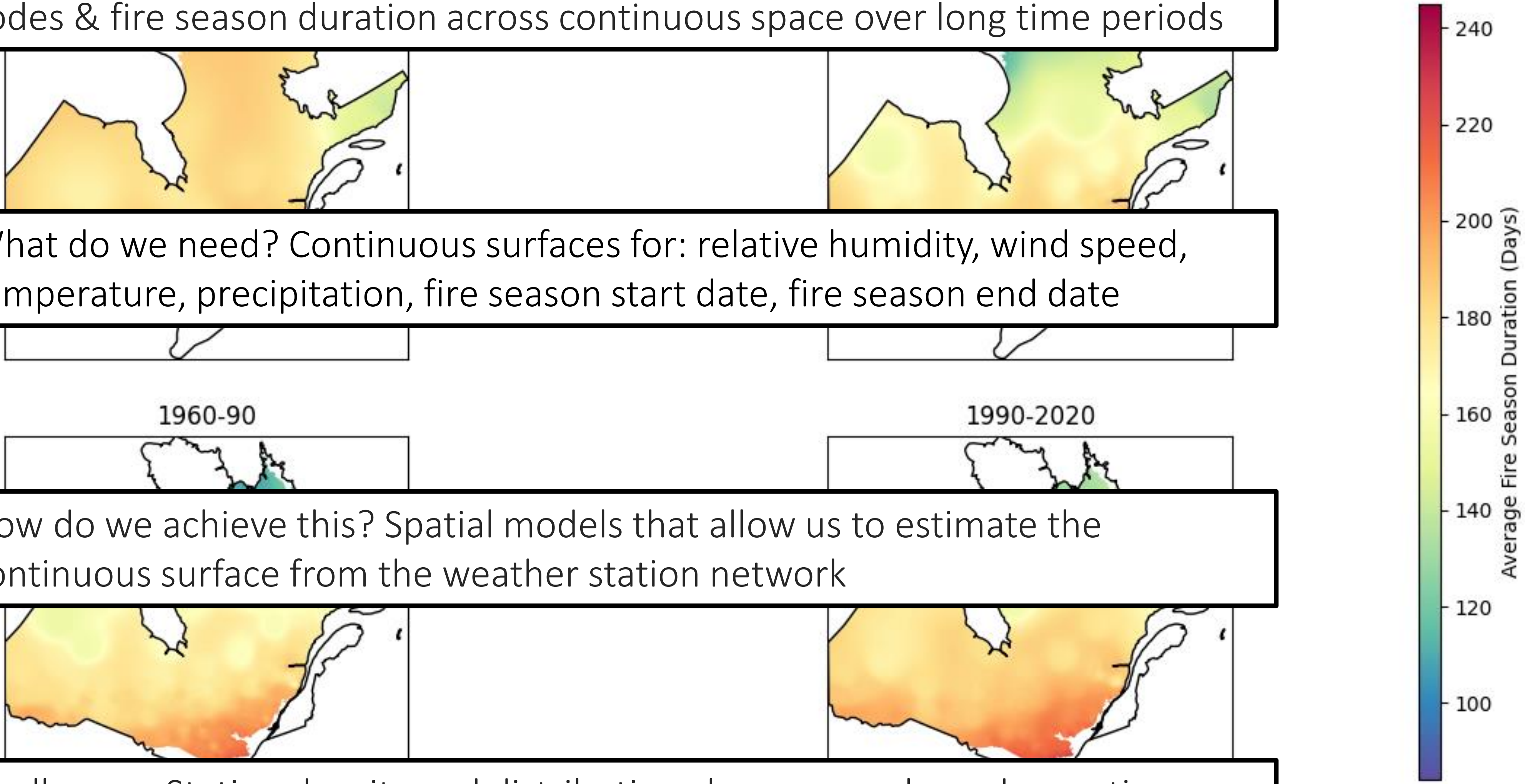
Cross-validation strategies for automatic selection of spatial models for forest meteorological applications

Clara Risk, November 17



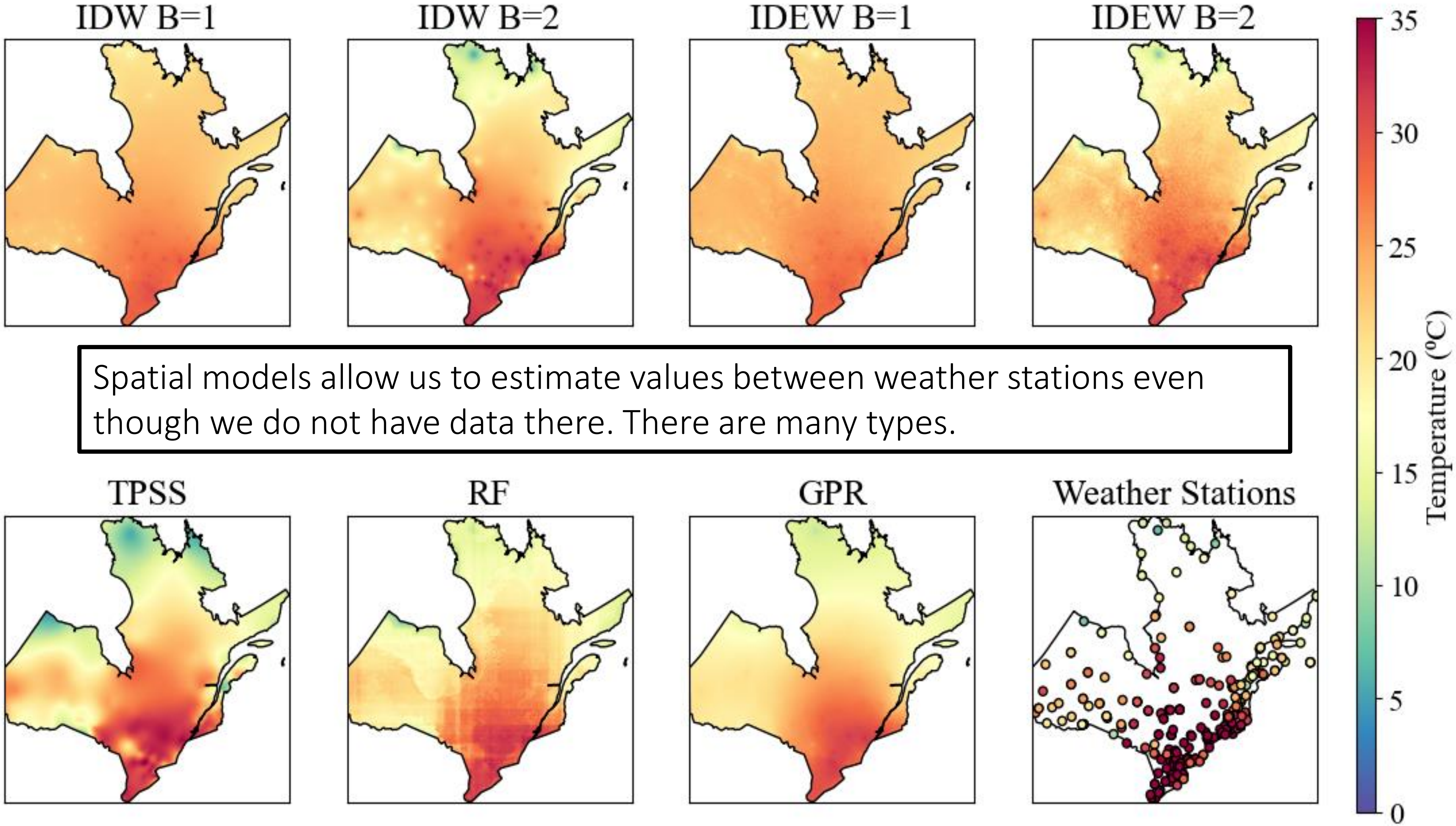
Overall Objective: Calculate the Canadian Forest Fire Weather Index System codes & fire season duration across continuous space over long time periods

What do we need? Continuous surfaces for: relative humidity, wind speed, temperature, precipitation, fire season start date, fire season end date



How do we achieve this? Spatial models that allow us to estimate the continuous surface from the weather station network

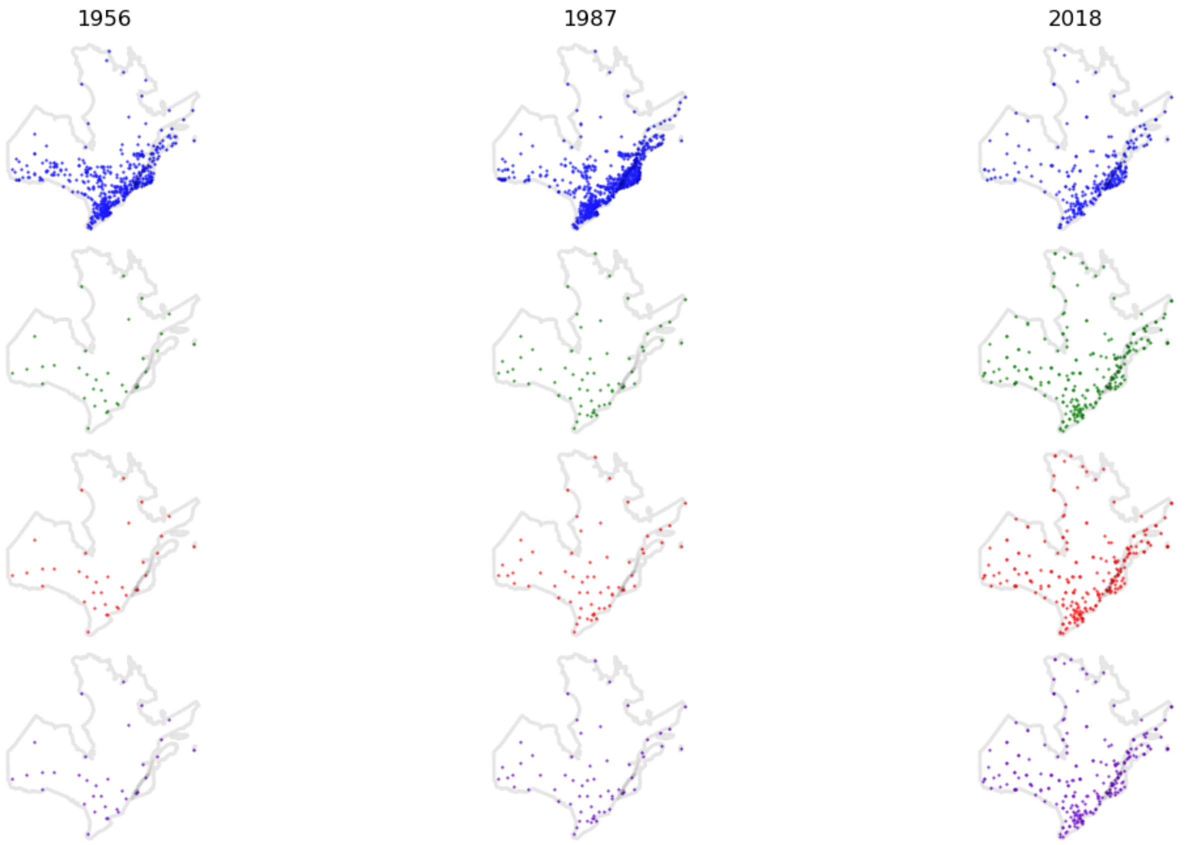
Challenges: Station density and distribution changes yearly, and sometimes daily or even hourly (if there is equipment failure)

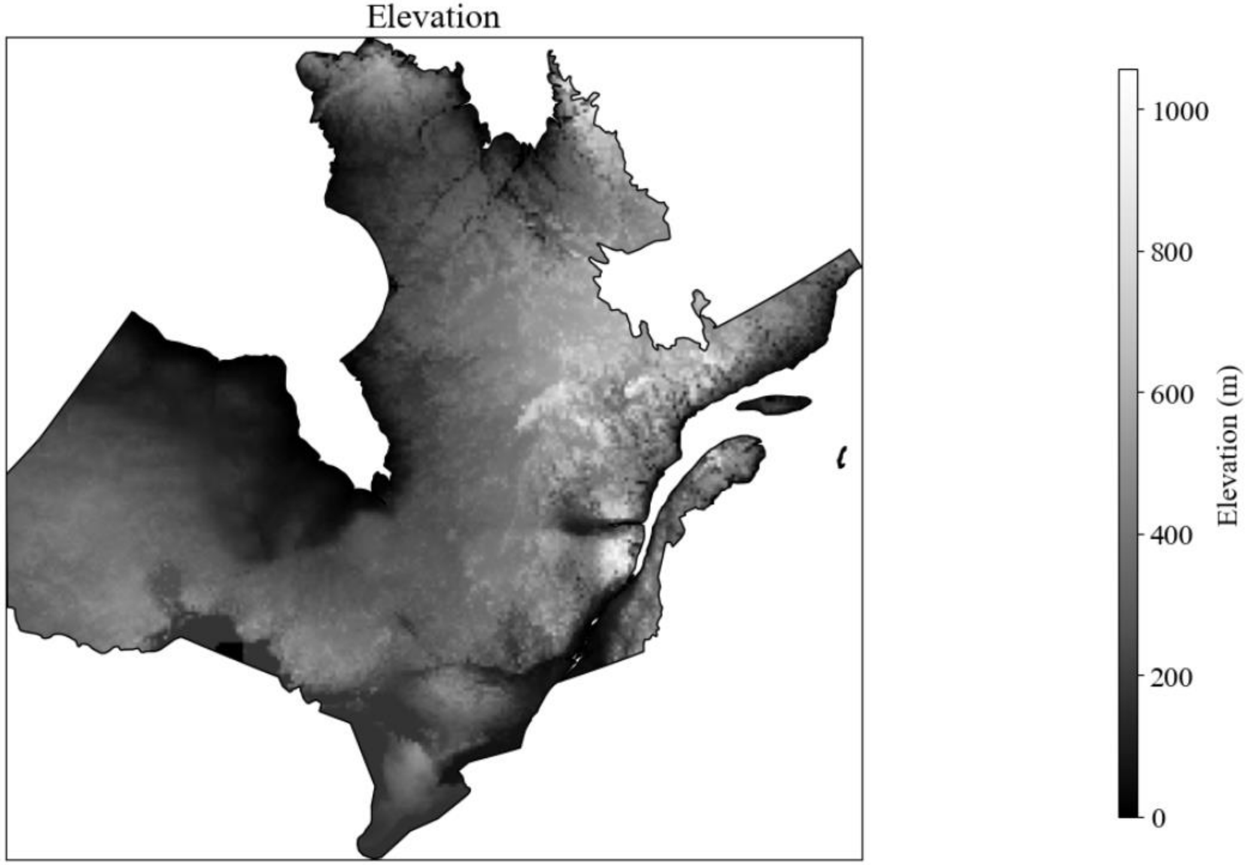


Spatial models allow us to estimate values between weather stations even though we do not have data there. There are many types.

FIG. 3. Surfaces produced for temperature for July 1, 2018 13:00 DST.

What data do we need? (A) Historical weather station data from Québec and Ontario & (B) Elevation

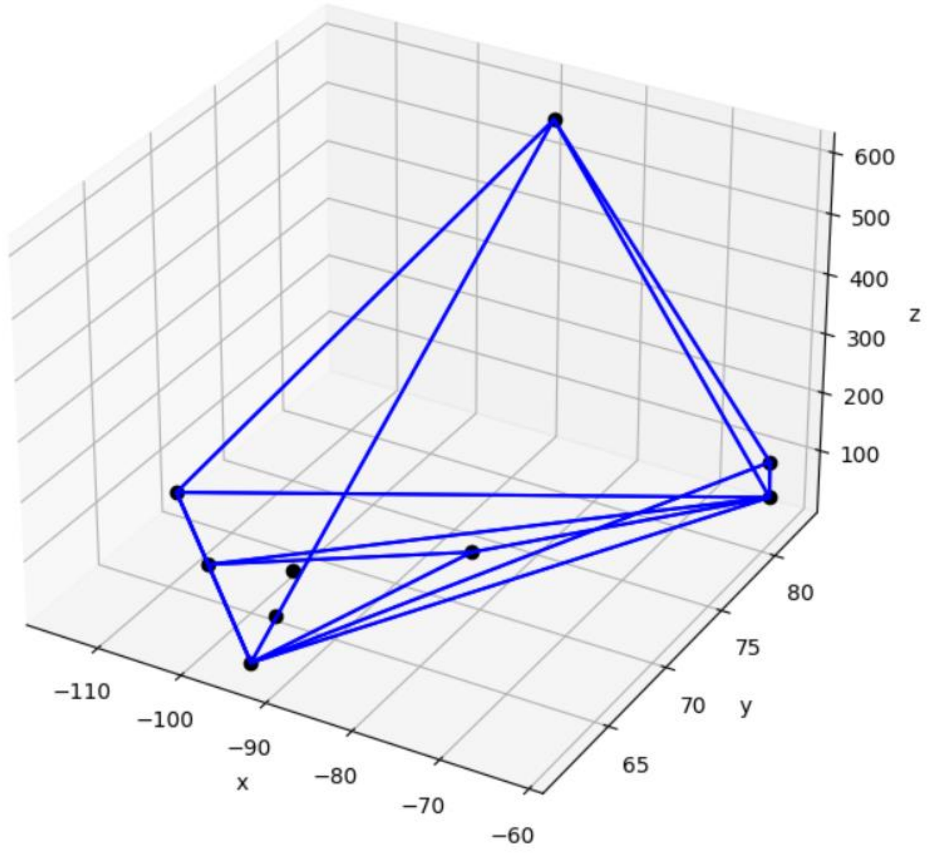




Spatial models perform two mathematical operations: spatial interpolation & spatial extrapolation

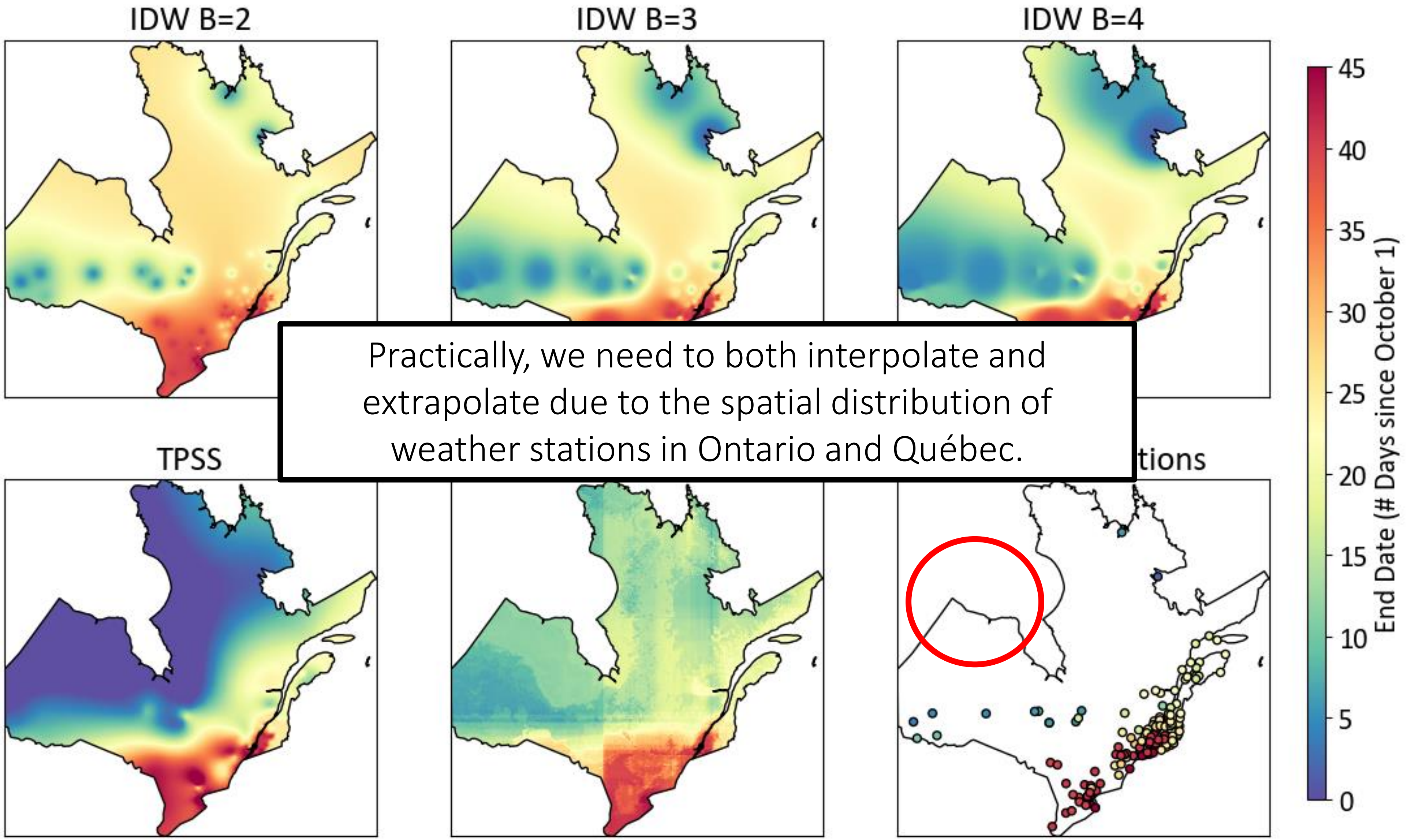
What is a convex hull?

What are spatial interpolation and extrapolation in meteorology?



Spatial interpolation → estimates generated inside the convex hull of the weather station network (Jain & Flannigan, 2017)

Spatial extrapolation → the estimates generated outside



Practically, we need to both interpolate and extrapolate due to the spatial distribution of weather stations in Ontario and Québec.

How do we evaluate the spatial models? Which one is the best?

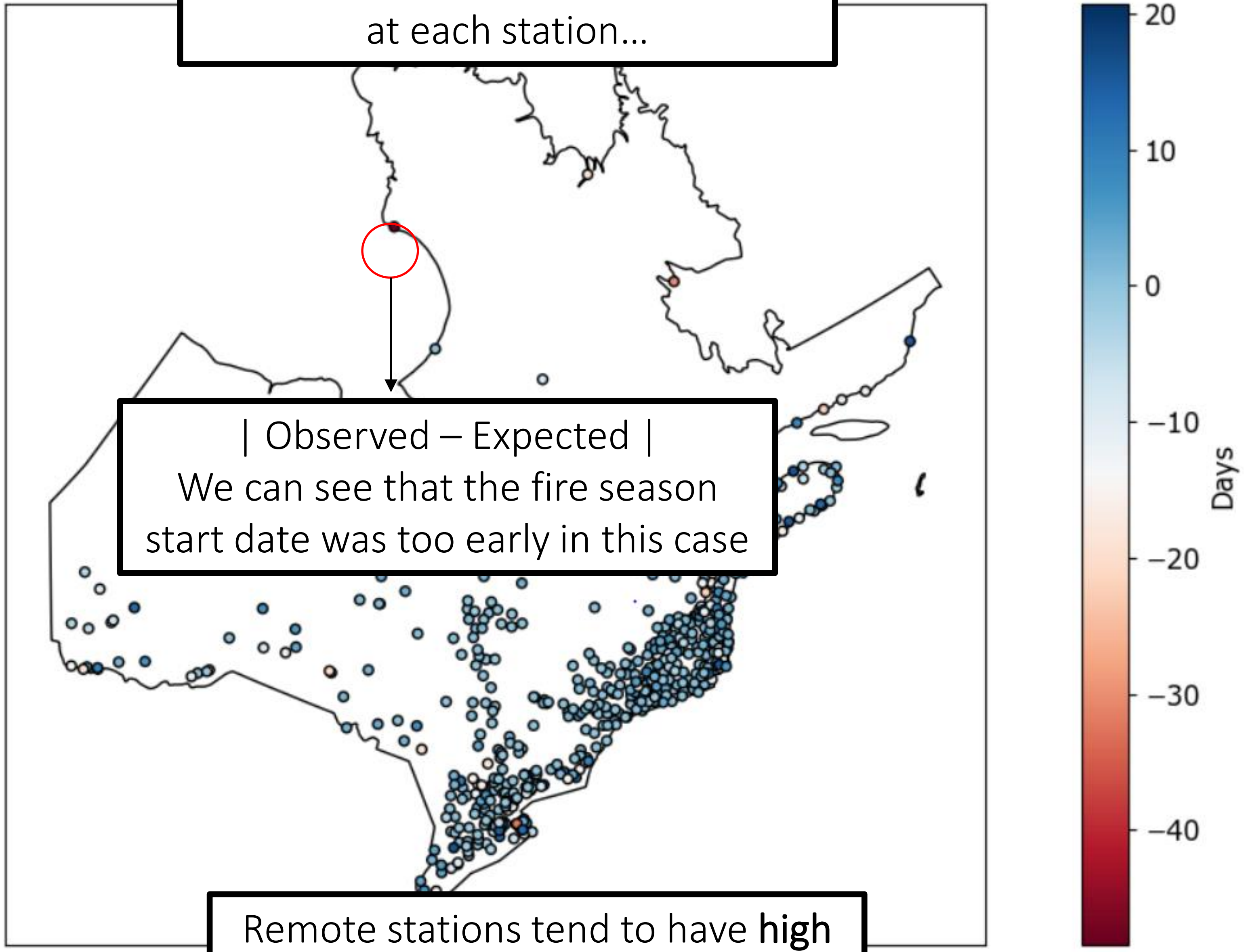
We evaluate with cross-validation

This involves progressively omitting weather station(s) from the spatial model and then comparing the observed versus expected results

The most common type in meteorology is leave-one-out cross-validation

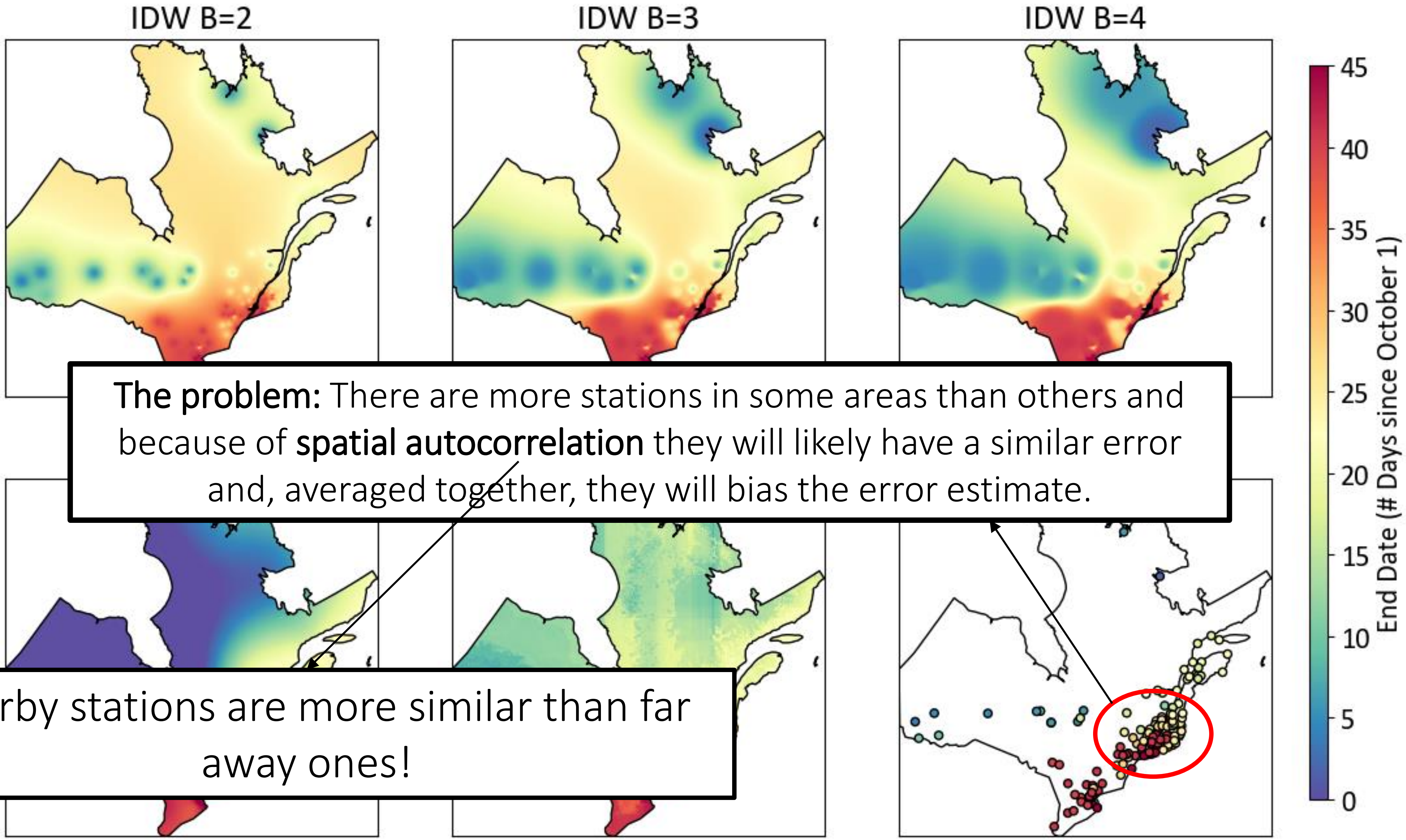
We progressively omit each weather station from the procedure, then calculate the average error for the network

We can generate an error estimate at each station…



| Observed – Expected |
We can see that the fire season start date was too early in this case

Remote stations tend to have **high leverage**



The problem: There are more stations in some areas than others and because of **spatial autocorrelation** they will likely have a similar error and, averaged together, they will bias the error estimate.

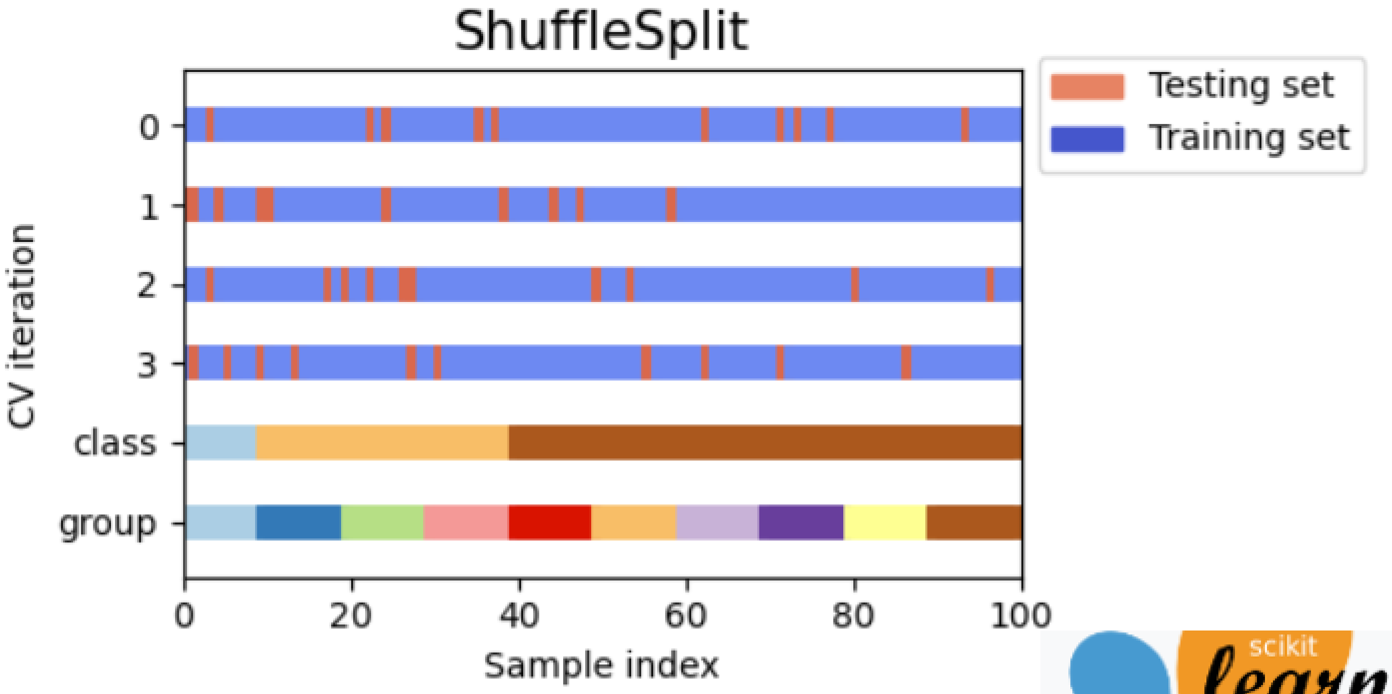
Nearby stations are more similar than far away ones!

Shuffle-Split Cross-Validation

Involves holding back a randomly selected (without replacement) group of weather stations a certain number of times

The error estimate, although still biased, is likely a truer estimate than that generated by the LOOCV procedure (Little et al., 2017)

We expect that the shuffle-split approach may not accurately reflect the true error of the surface because it does not account for the underlying spatial dependence of the error (Hutchinson et al., 2009; Roberts et al., 2017)



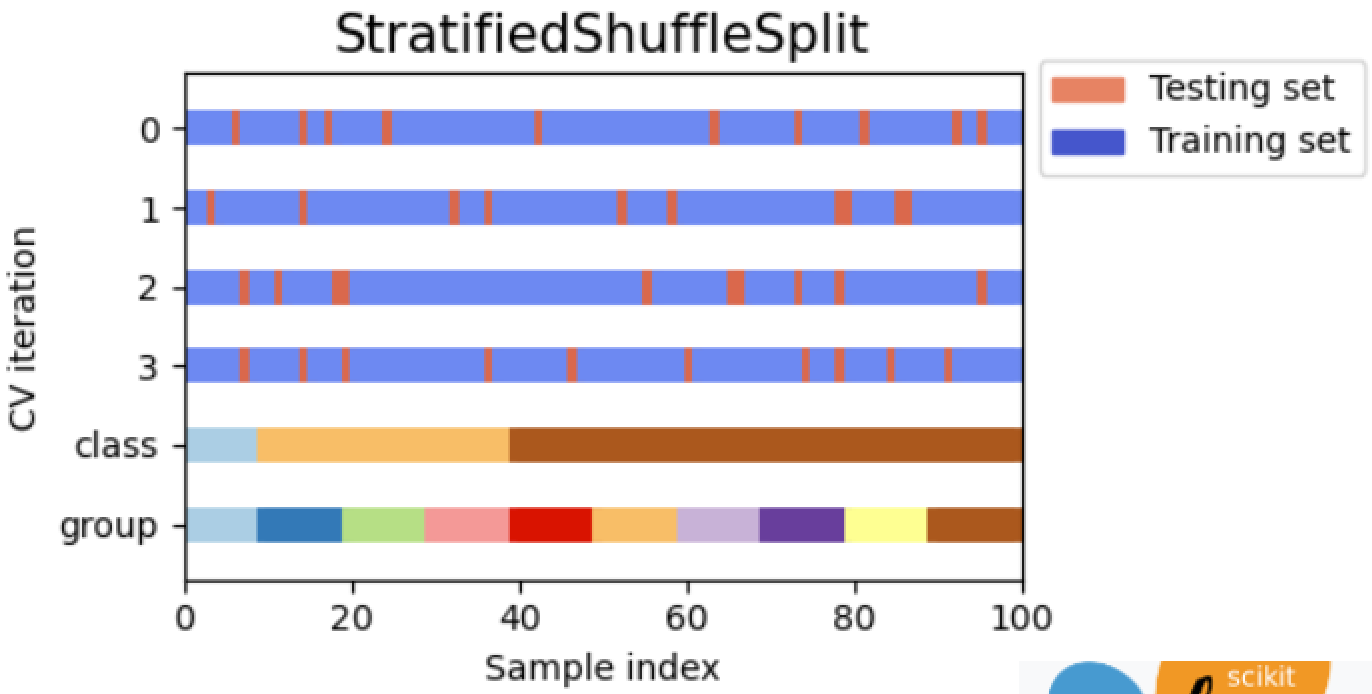
Spatial Bootstrap (Stratified Shuffle-Split)

Make groups of stations

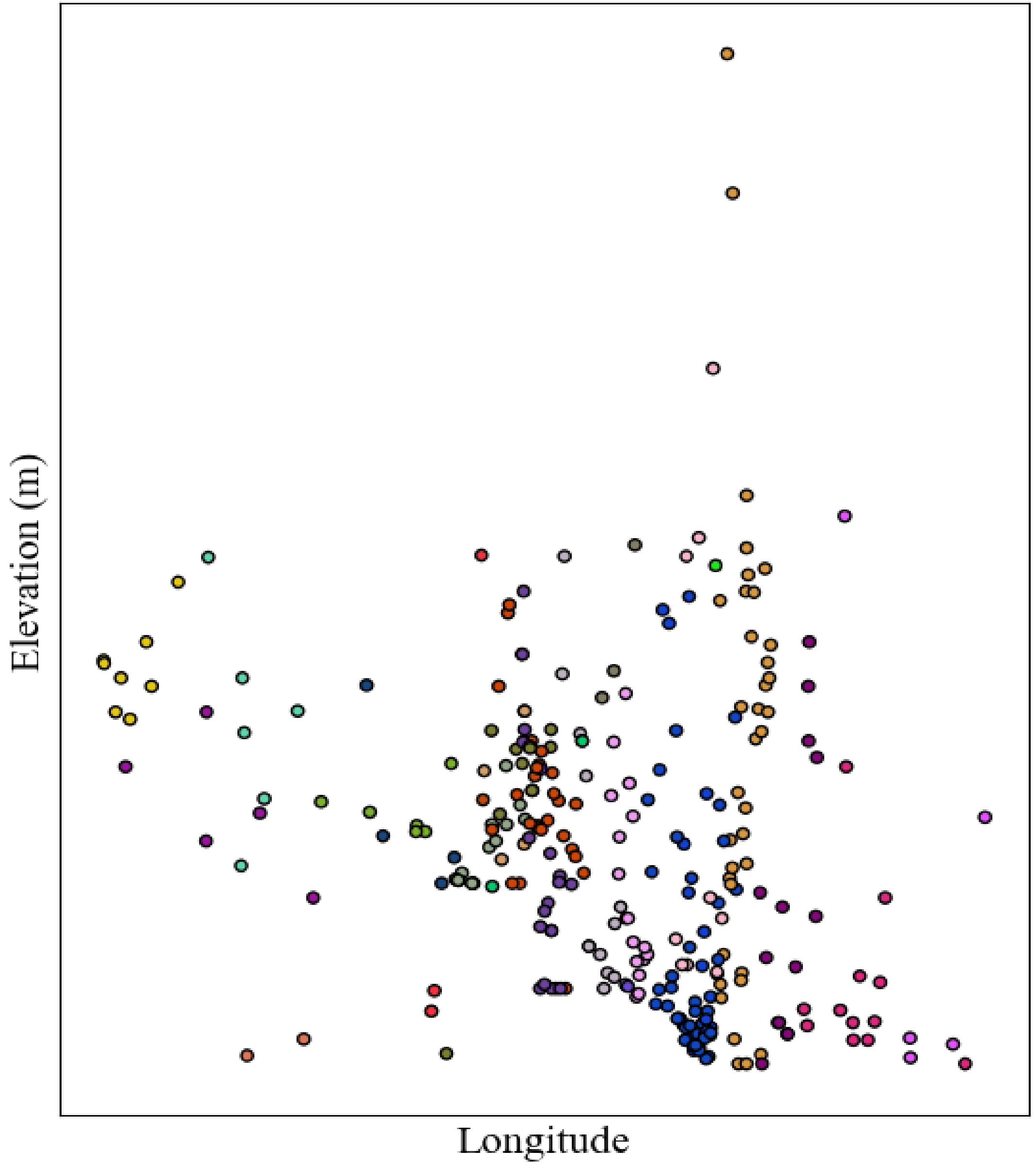
Randomly select a station from each group

Ensures that the southern stations do not dominate the error estimate

But how do we define the groups? How do we decide the number of groups?



We will use clustering to create the groups



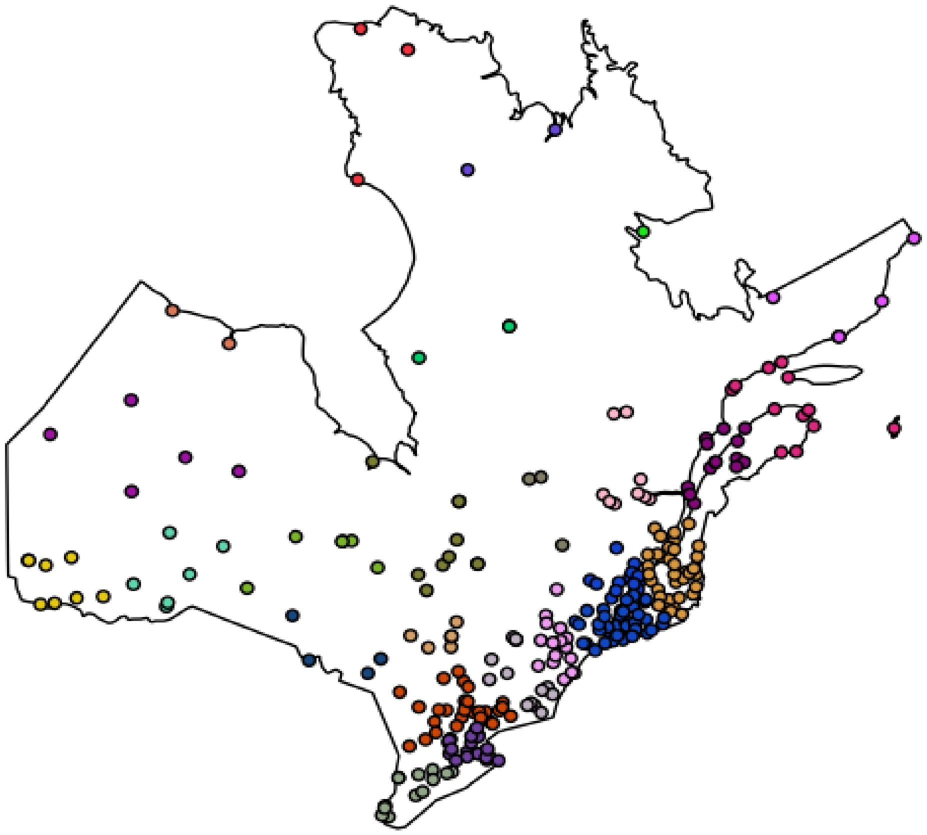
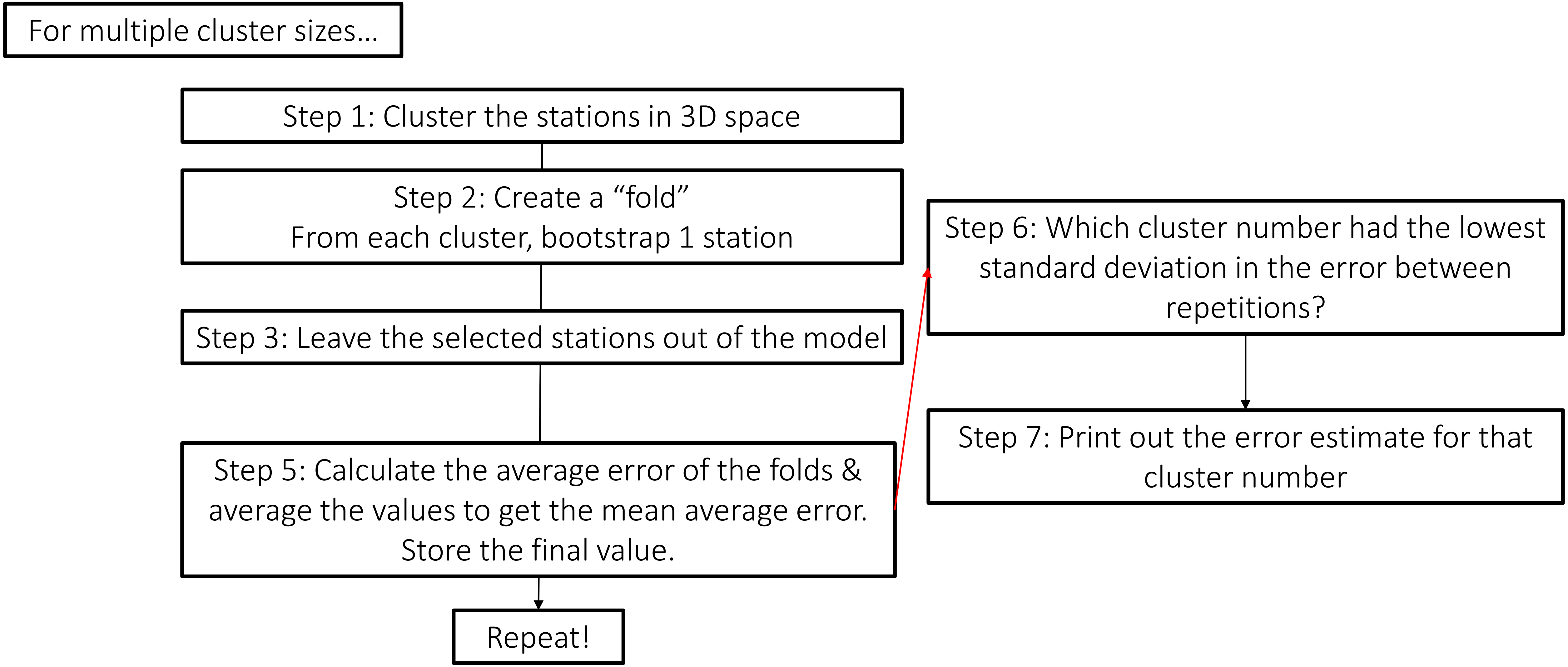
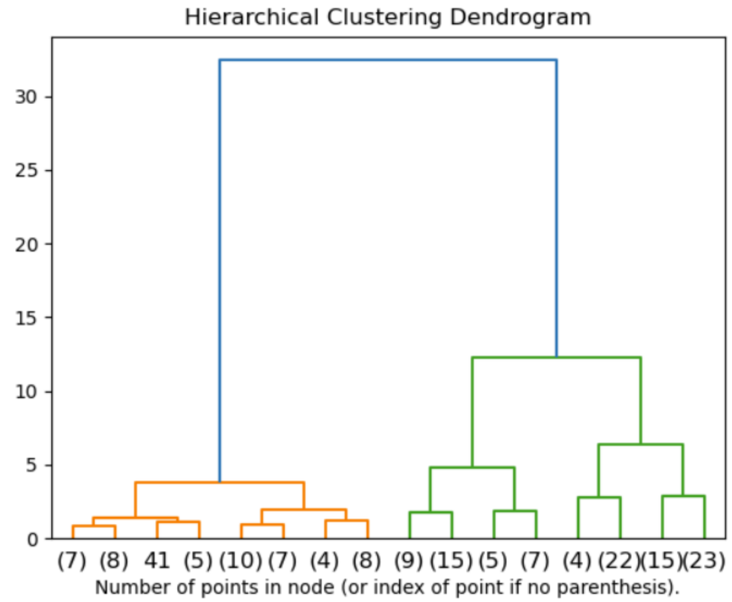


FIG. 1. Results of spatial clustering (25 clusters) for precipitation for July 1, 2018.

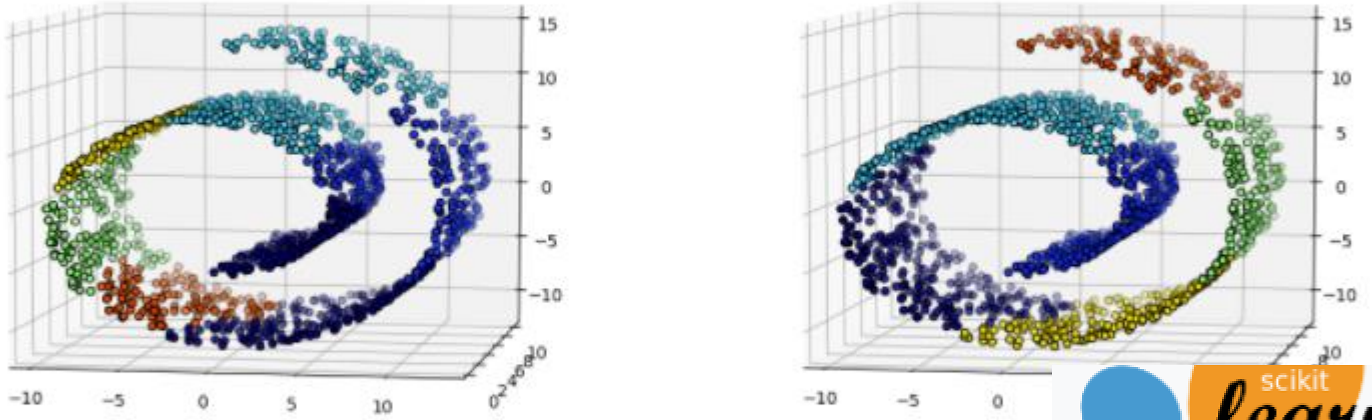

For multiple cluster sizes…
Step 1: Cluster the stations in 3D space
Step 2: Create a "fold"
From each cluster, bootstrap 1 station
Step 3: Leave the selected stations out of the model
Step 5: Calculate the average error of the folds & average the values to get the mean average error. Store the final value.
Repeat!
Step 6: Which cluster number had the lowest standard deviation in the error between repetitions?
Step 7: Print out the error estimate for that cluster number

Step 1: Agglomerative clustering with Ward's linkage using latitude, longitude, and elevation

Merge clusters with smallest sum of squared differences between them at each split (Bhandari & Pahwa, 2020)



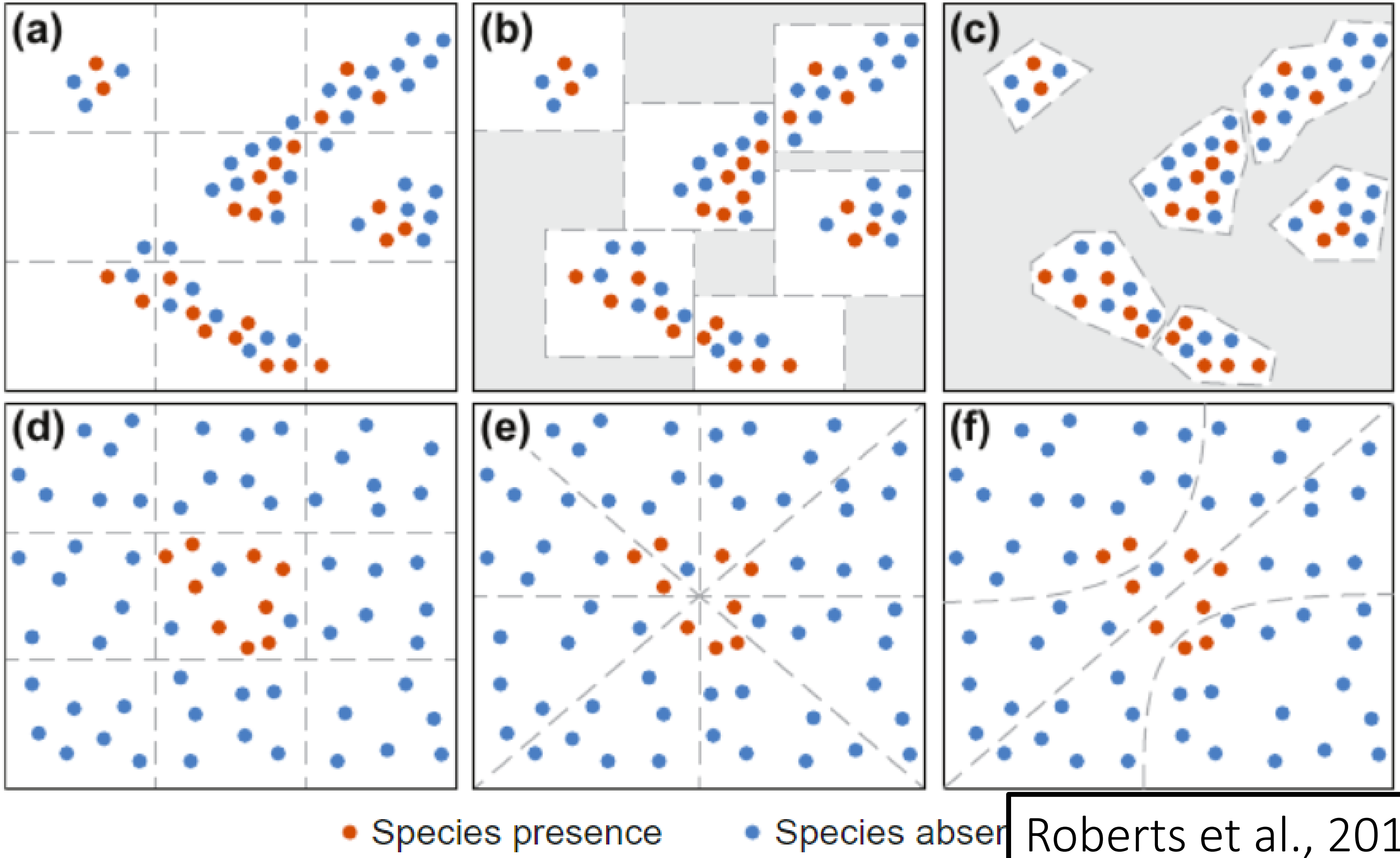
Each station begins as a single cluster and then new clusters are formed by calculating similarity (Singh et al., 2011)



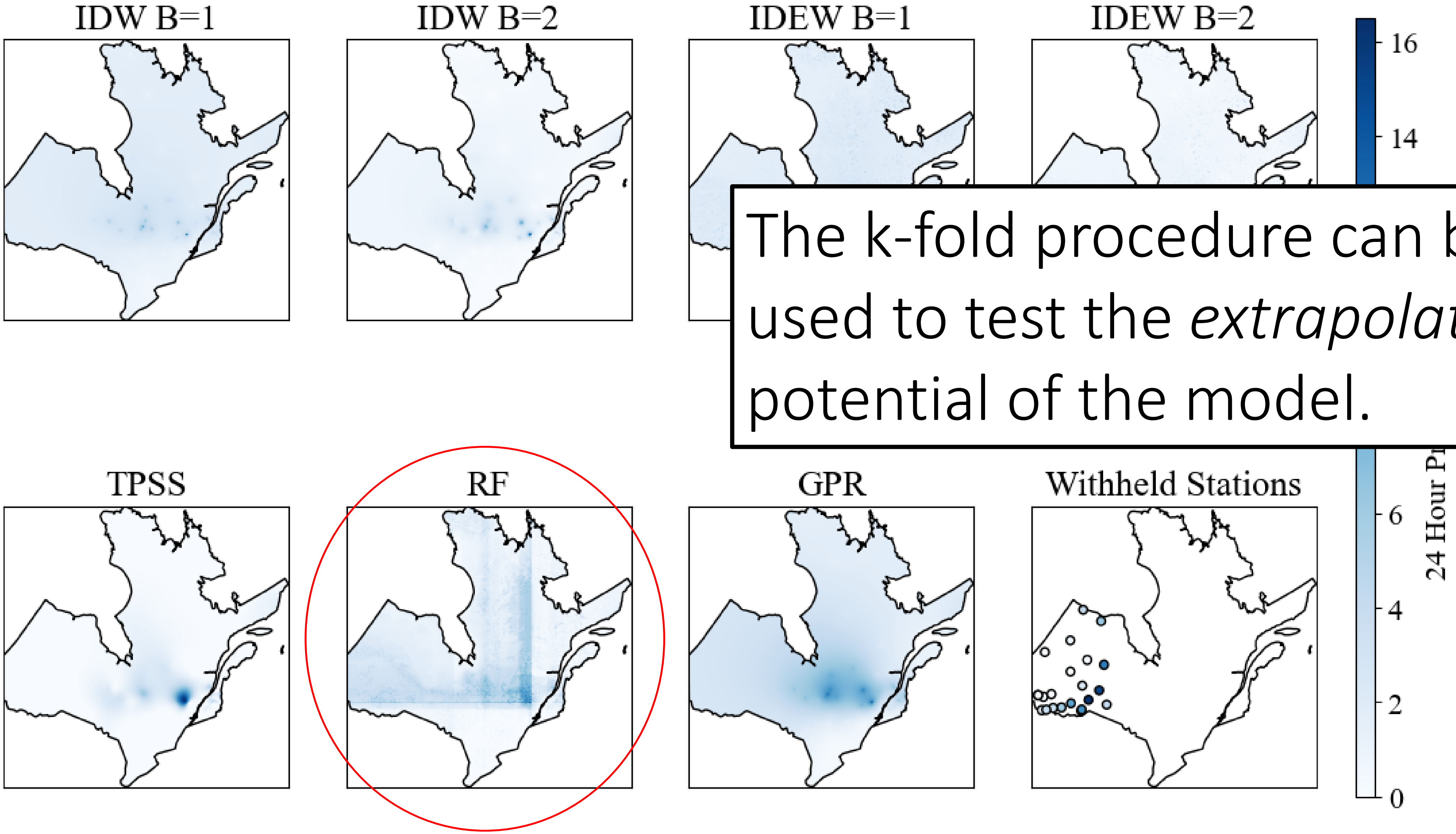
Implementation of a connectivity matrix to ensure we only merge neighboring clusters at each split point in the dendrogram (Thirion et al., 2014)

Spatial k-Fold Cross-Validation

Leave out blocks of samples progressively



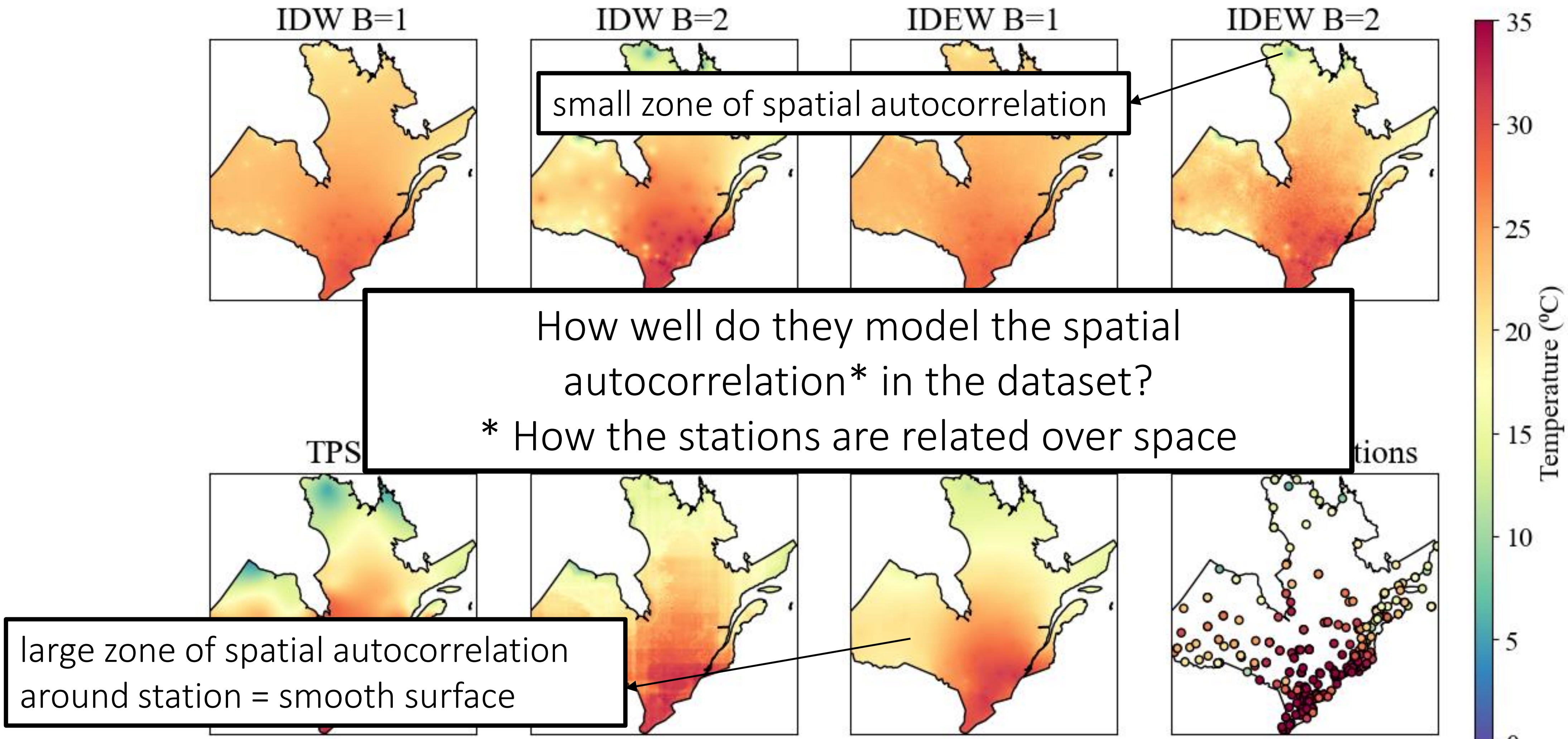
Roberts et al., 2017



The k-fold procedure can be used to test the *extrapolation* potential of the model.

FIG. 7. Surfaces produced for precipitation when all stations in northwestern Ontario were left out for July 1, 2018.

Spatial models in depth



small zone of spatial autocorrelation

How well do they model the spatial autocorrelation* in the dataset?
* How the stations are related over space

large zone of spatial autocorrelation around station = smooth surface

FIG. 3. Surfaces produced for temperature for July 1, 2018 13:00 DST.

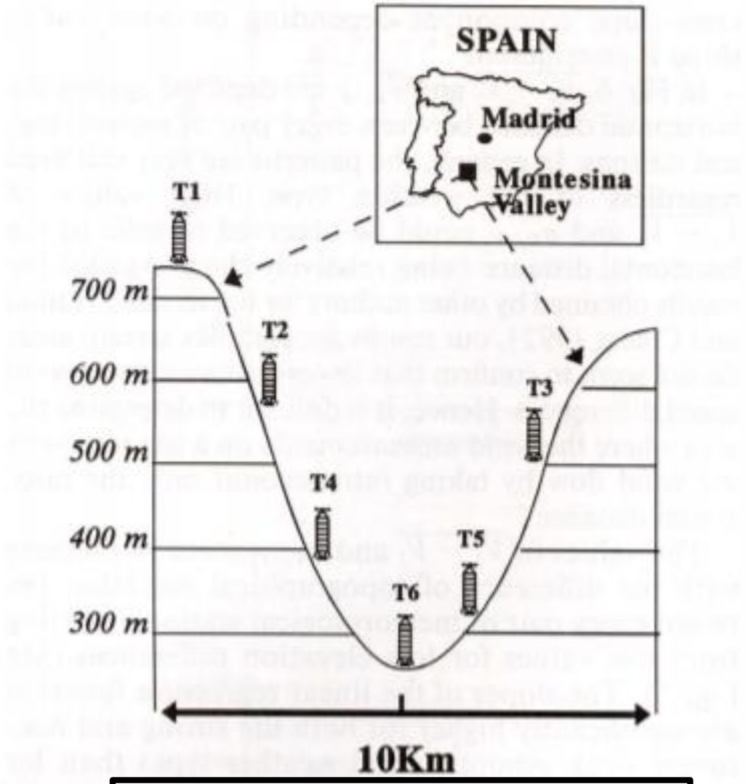
Spatial model: Inverse distance elevation weighting (IDEW)

$$W = 0.8(1/D^{-\beta}) * 0.2(1/H^{-\beta})$$

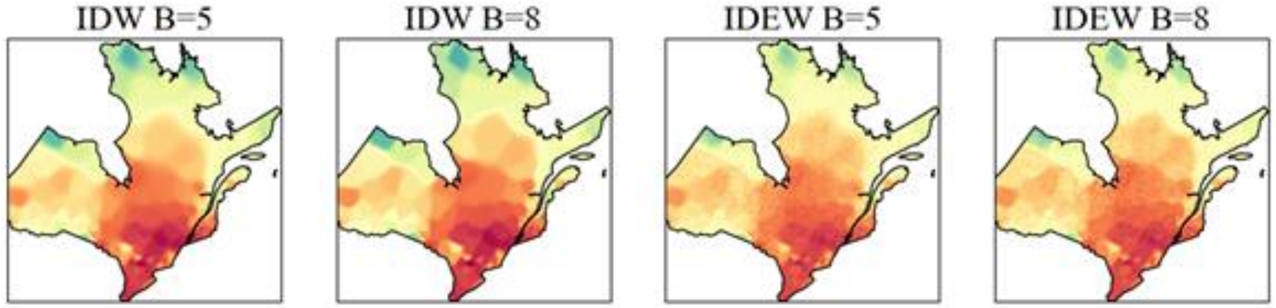
Weight of station at location

Exponent controlling weight of farther away points

IDEW assigns weights to different weather stations according to their distance from the point being estimated



Palomino & Martin, 1995



Advantage: simple, computationally efficient
Disadvantage: does not produce a smooth surface, cannot extrapolate well

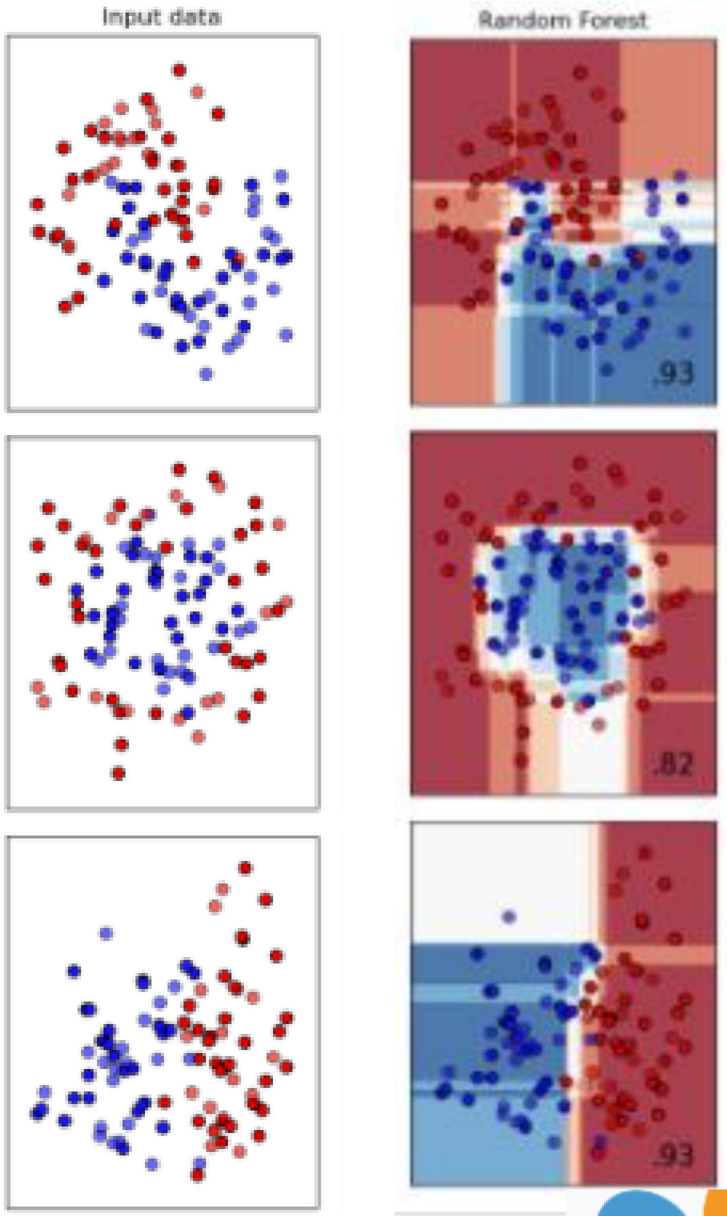
Assumptions? The modelled spatial autocorrelation between stations is uniform across the dataset (controlled by β)

Spatial model: Random Forests (RF)

RF is a machine learning method that involves creating many decision trees by randomly selecting samples from a dataset with replacement (Breiman, 2001)

Advantage: superior for extrapolation
Disadvantage: it usually results in banding on the continuous surface, moderately computationally intensive

Assumptions? Not many! But it does assume that the values of the weather stations are representative of the actual surface (i.e., not an extreme outlier)



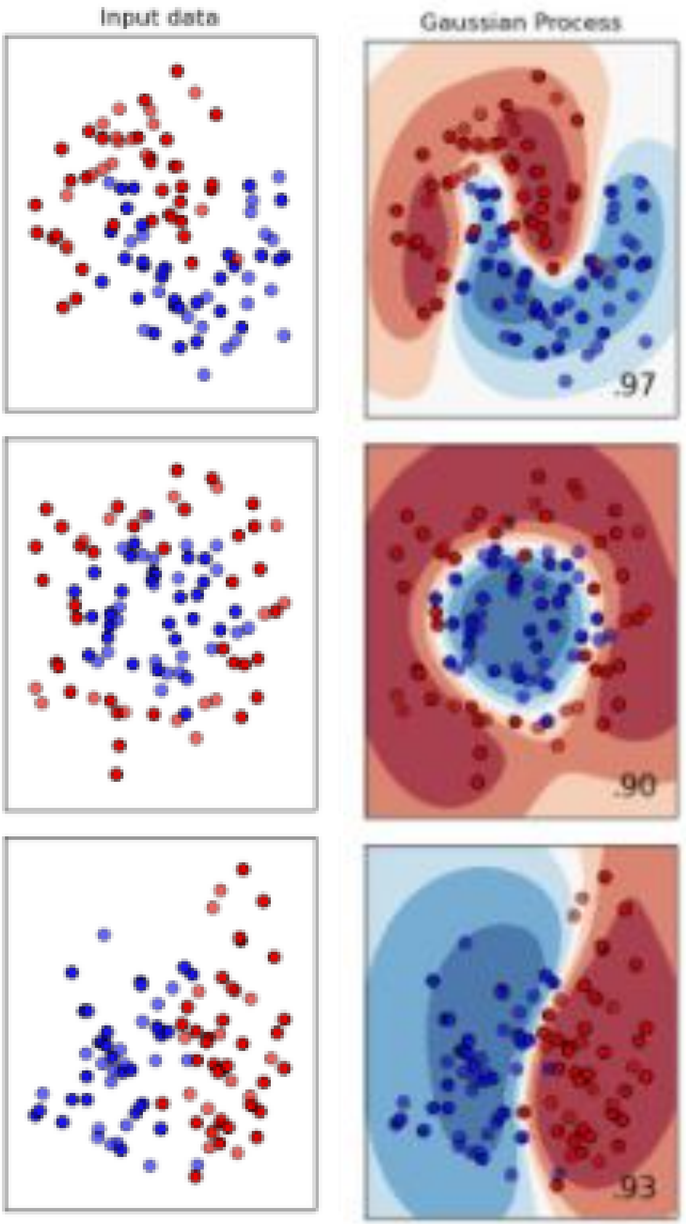
Spatial model: Gaussian Process Regression (GPR)

With linear regression, we assume that the independent and dependent variables are related uniformly over space… GPR allows that relationship to vary (Rasmussen & Williams, 2006)

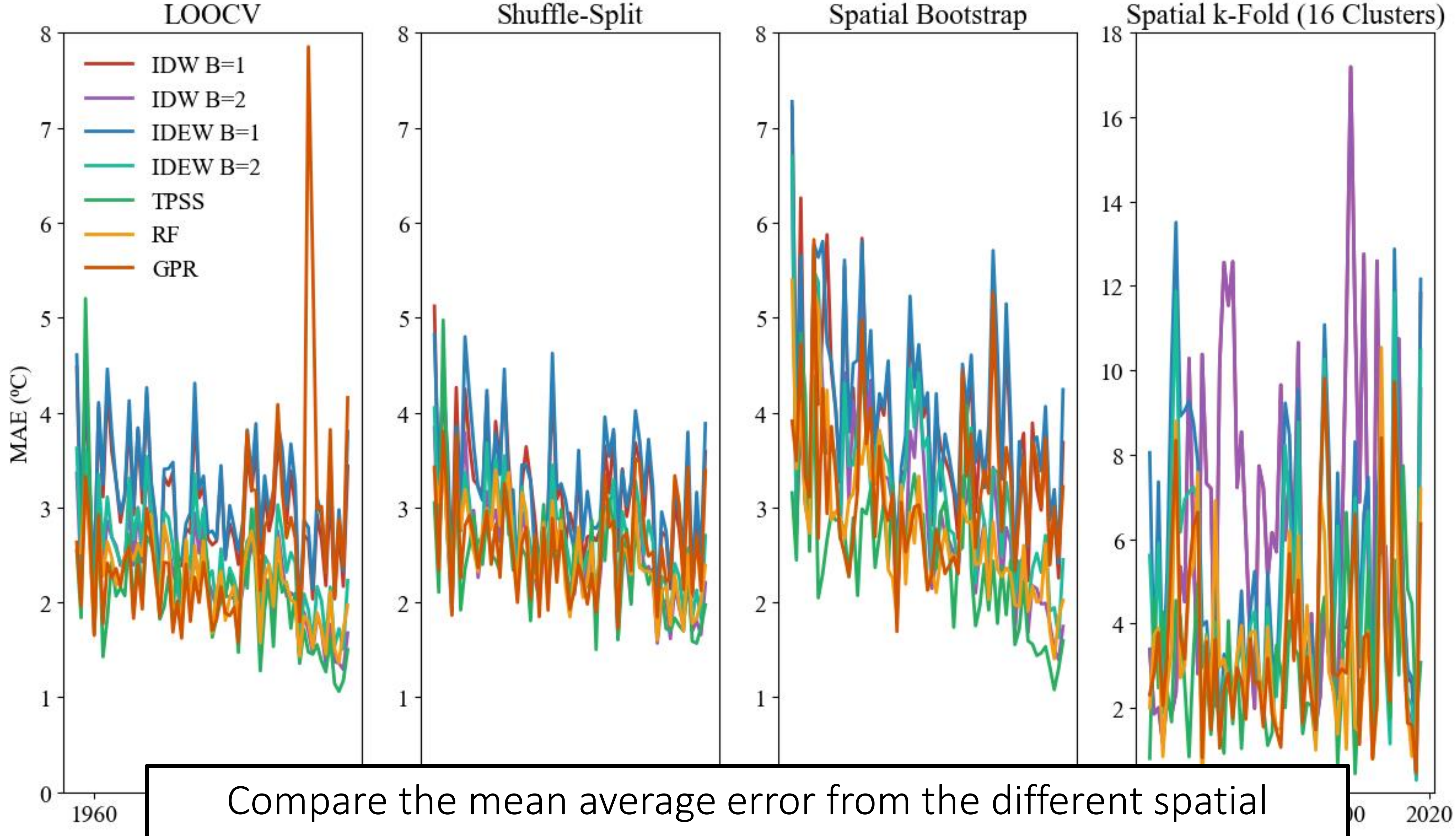
The procedure uses the known values to fit many functions that are randomly sampled from the Gaussian (normal) distribution (Rasmussen & Williams, 2006)

Advantage: smooth surface characteristic of weather variables
Disadvantage: computationally intensive, covariance function may not always fit the dataset on a certain day

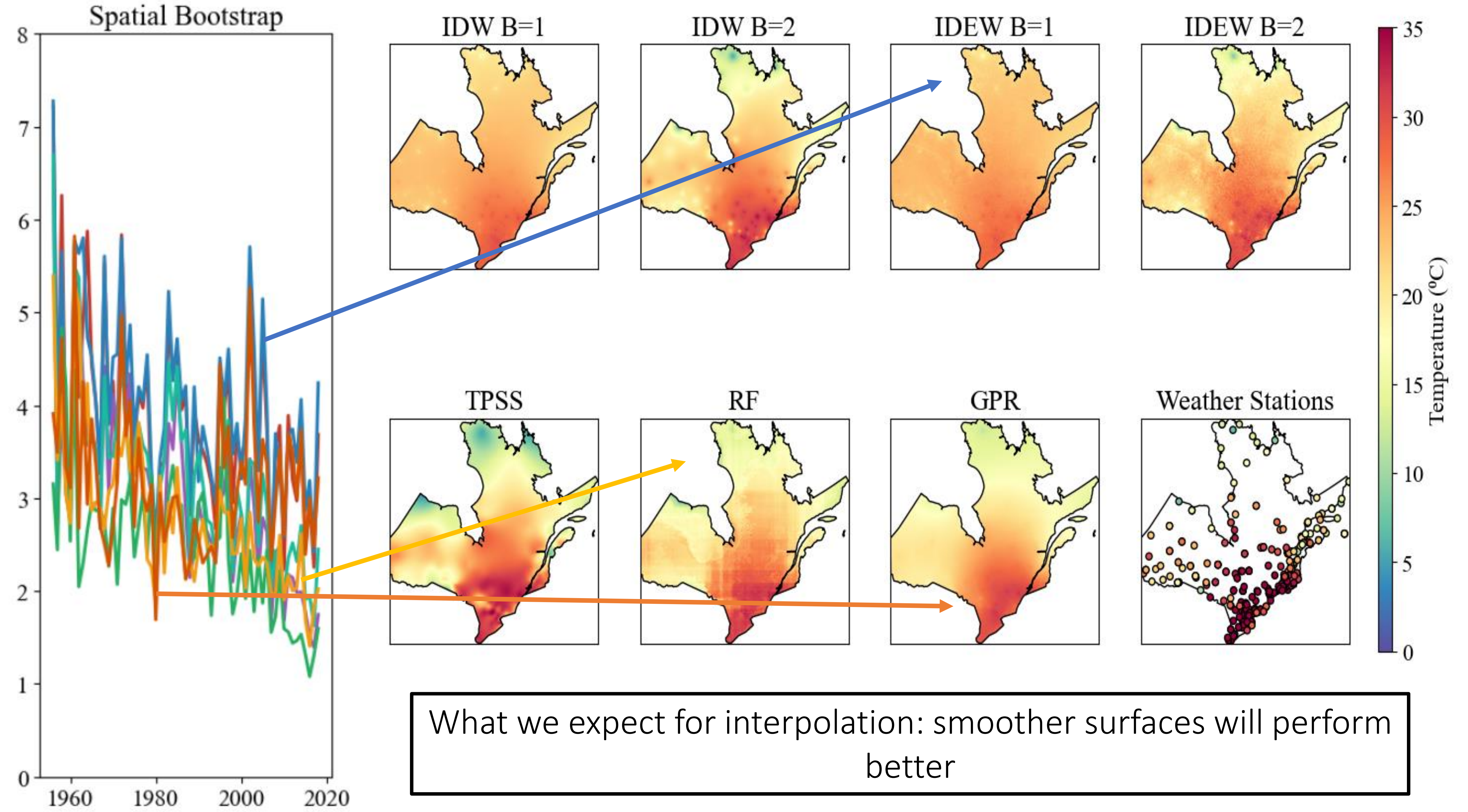
Assumptions? Assumes the covariance function you select describes the spatial autocorrelation across the surface. If too strict, the range of values we will consider at a certain point will be too narrow.



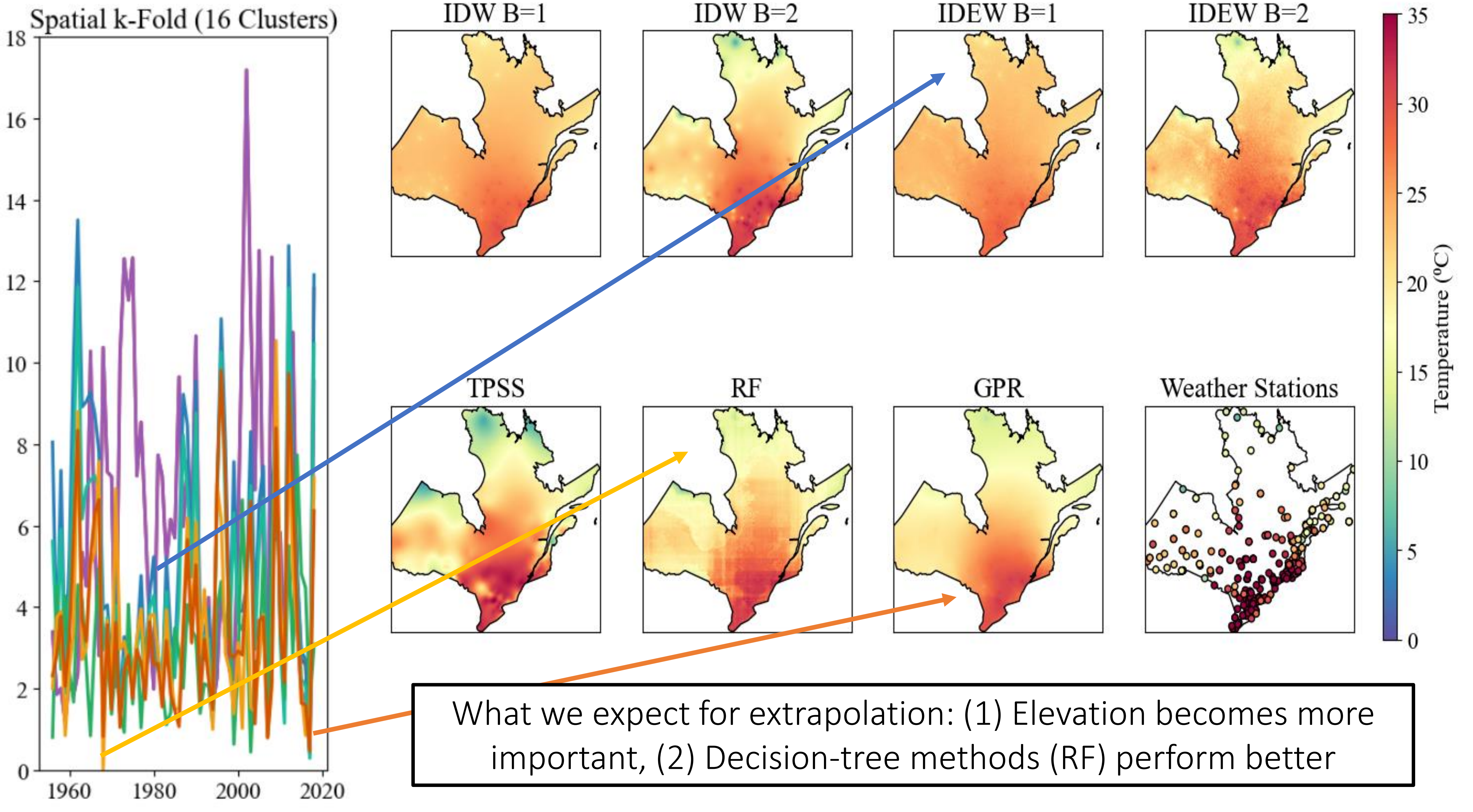
The comparison



Compare the mean average error from the different spatial models using the four types of cross-validation



What we expect for interpolation: smoother surfaces will perform better



What we expect for extrapolation: (1) Elevation becomes more important, (2) Decision-tree methods (RF) perform better

References

Danielson, J. J., & Gesh, D. B. (2011). Global multi-resolution terrain elevation data 2010 (GMTED2010): U.S. Geological Survey Open-File Report 2011–1073. In *U.S. Geological Survey Open-File Report 2011-1073* (Vol. 2010).

Jain, P., & Flannigan, M. D. (2017). Comparison of methods for spatial interpolation of fire weather in Alberta, Canada. *Canadian Journal of Forest Research*, *47*(12), 1646–1658. https://doi.org/10.1139/cjfr-2017-0101

Hutchinson, Michael F., McKenney, D. W., Lawrence, K., Pedlar, J. H., Hopkinson, R. F., Milewska, E., & Papadopol, P. (2009). Development and testing of Canada-wide interpolated spatial models of daily minimum-maximum temperature and precipitation for 1961-2003. *Journal of Applied Meteorology and Climatology*, *48*(4), 725–741. https://doi.org/10.1175/2008JAMC1979.1

Little, M. A., Varoquaux, G., Saeb, S., Lonini, L., Jayaraman, A., Mohr, D. C., & Kording, K. P. (2017). Using and understanding cross-validation strategies. Perspectives on Saeb et al. In *GigaScience* (Vol. 6, Issue 5, pp. 1–6). https://doi.org/10.1093/gigascience/gix020

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., & Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, *40*(8), 913–929. https://doi.org/10.1111/ecog.02881

Dale, M. R. T., & Fortin, M. J. (2002). Spatial autocorrelation and statistical tests in ecology. *Ecoscience*, *9*(2), 162–167. https://doi.org/10.1080/11956860.2002.11682702

Fortin, M. J., & Jacquez, G. M. (2000). Randomization tests and spatially auto-correlated data. *Bulletin of the Ecological Society of America*, *81*(3), 201–205. http://www.jstor.org/stable/20168439

Bhandari, N., & Pahwa, P. (2020). Evaluating performance of agglomerative clustering for extended NMF. *Journal of Statistics and Management Systems*, 1–12. https://doi.org/10.1080/09720510.2020.1799507

Singh, W., Hjorleifsson, E., & Stefansson, G. (2011). Robustness of fish assemblages derived from three hierarchical agglomerative clustering algorithms performed on Icelandic groundfish survey data. *ICES Journal of Marine Science*, *68*(1), 189–200. https://doi.org/10.1093/icesjms/fsq144

Thirion, B., Varoquaux, G., Dohmatob, E., & Poline, J. B. (2014). Which fMRI clustering gives good brain parcellations? *Frontiers in Neuroscience*, *8 JUL*. https://doi.org/10.3389/fnins.2014.00167

Palomino, I., & Martin, F. (1995). A simple method for spatial interpolation of the wind in complex terrain. *Journal of Applied Meteorology*, *34*(7), 1678–1693. https://doi.org/10.1175/1520-0450-34.7.1678

da Silva Júnior, J. C., Medeiros, V., Garrozi, C., Montenegro, A., & Gonçalves, G. E. (2019). Random forest techniques for spatial interpolation of evapotranspiration data from Brazilian's Northeast. *Computers and Electronics in Agriculture*, *166*, 105017. https://doi.org/10.1016/j.compag.2019.105017

Guan, H., Li, J., Chapman, M., Deng, F., Ji, Z., & Yang, X. (2013). Integration of orthoimagery and lidar data for object-based urban thematic mapping using random forests. *International Journal of Remote Sensing*, *34*(14), 5166–5186. https://doi.org/10.1080/01431161.2013.788261

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Rasmussen, C. E., & Williams, C. K. I. (2006). Gaussian processes for machine learning. 2006. In *The MIT Press, Cambridge, MA, USA* (Vol. 38, Issue 2).

scikit-learn developers. (2019). sklearn.gaussian_process.kernels.RationalQuadratic. Retrieved September 26, 2020, from https://scikit-learn.org/stable/modules/generated/sklearn.gaussian_process.kernels.RationalQuadratic.html#sklearn.gaussian_process.kernels.RationalQuadratic

scikit-learn developers. (2019). Nearest Neighbor Classification. Retrieved November 4, 2020, from https://scikit-learn.org/stable/auto_examples/neighbors/plot_classification.html#sphx-glr-auto-examples-neighbors-plot-classification-py

scikit-learn developers. (2019). Hierarchical clustering: structured vs unstructured ward. Retrieved November 4, 2020, from https://scikit-learn.org/stable/auto_examples/cluster/plot_ward_structured_vs_unstructured.html#sphx-glr-auto-examples-cluster-plot-ward-structured-vs-unstructured-py

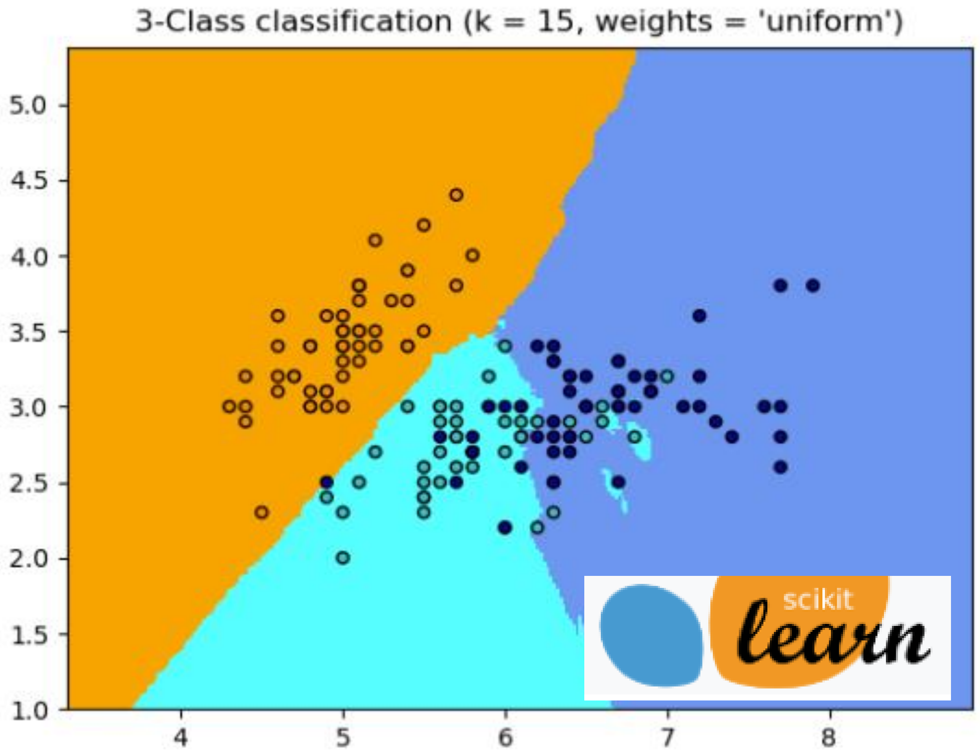
scikit-learn developers. (2019). Plot Hierarchical Clustering Dendrogram. Retrieved November 4, 2020, from https://scikit-learn.org/stable/auto_examples/cluster/plot_agglomerative_dendrogram.html#sphx-glr-auto-examples-cluster-plot-agglomerative-dendrogram-py

scikit-learn developers. (2019). Visualizing cross-validation behavior in scikit-learn. Retrieved November 4, 2020, from https://scikit-learn.org/stable/auto_examples/model_selection/plot_cv_indices.html#sphx-glr-auto-examples-model-selection-plot-cv-indices-py

Why are these spatial models important? Why do we need them?

People need to use weather & fire season duration as a covariate in their model. For example, if you have a seed source, you will want climate estimates for that exact location.

Why not just use the nearest weather station?



This method exists and is called Nearest Neighbour Interpolation (Thiessen/Voronoi diagrams). However, it is not very accurate in most cases for weather variables, because it results in abrupt transitions on the landscape that are not realistic.

Spatial model: Gaussian Process Regression (GPR)… the math!

$$m(x) = E[f(\mathrm{x})] = 0 \text{ or Constant}$$

Set the mean function to the mean value from the weather stations
E is the Expectation → the combination of the possible values generated from the random functions that we sample from the Gaussian distribution, or f(x) (Rasmussen & Williams, 2006)

$$k(x, x') = E[(f(\mathrm{x}) - m(x))(f(x') - m(x'))]$$

The mean function is used to create the covariance function, which describes how fast the spatial autocorrelation drops off as we move out farther from a station (scikit-learn developers, 2019)

$$f(x) = GP(m(x), k(x, x')$$

Combine into a single equation…

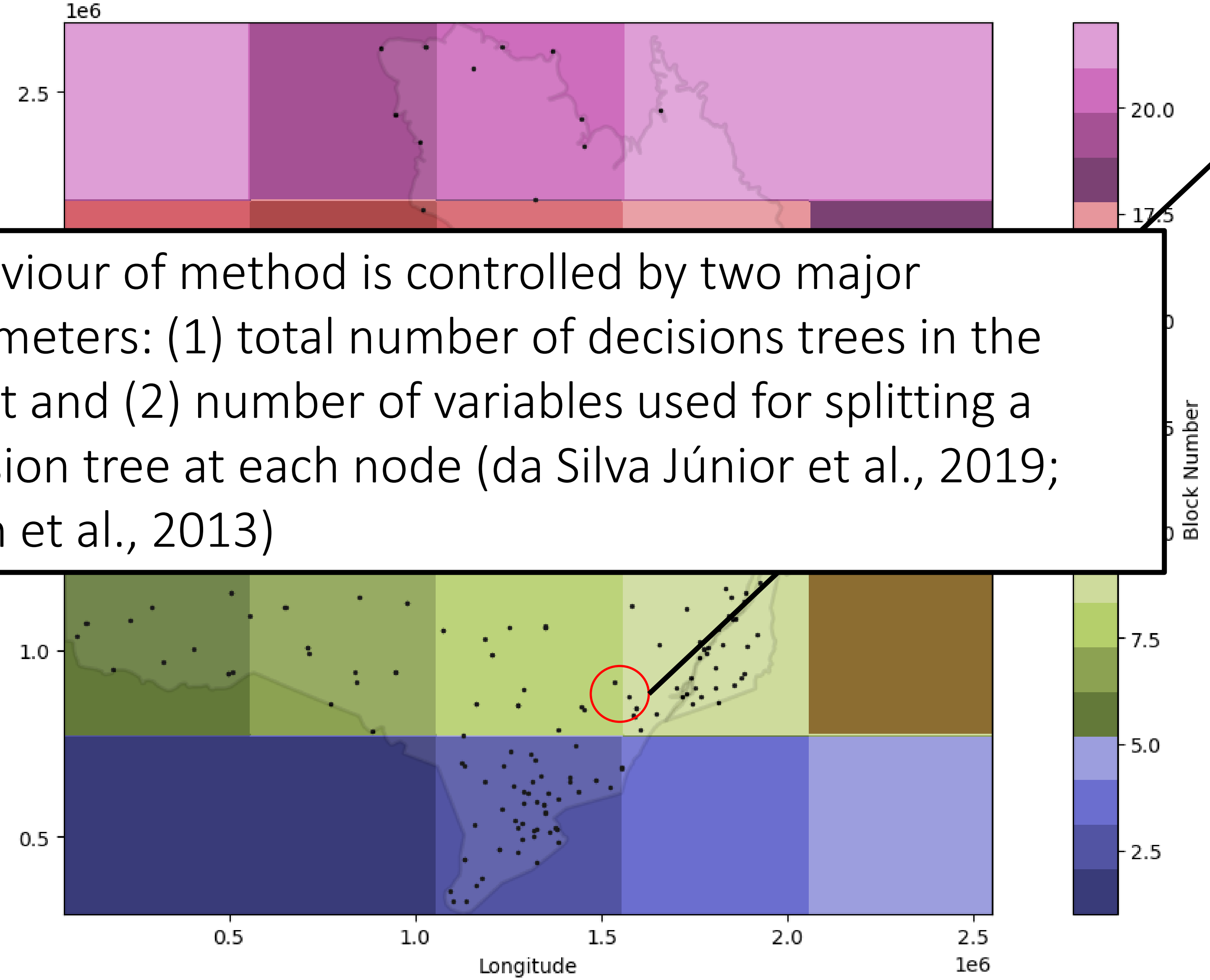
Select the covariance function – there are many! We selected the rational quadratic kernel (*a* controls spatial autocorrelation)

$$k(x, x') = \left(1 + \frac{d(x, x')^2}{2al^2}\right)^{-a}$$

What is it? The sum of an infinite series of radial basis functions (Rasmussen & Williams, 2006)

Why this one? Radial basis functions (such as thin plate splines) perform well in our study area for interpolation of temperature

Arbitrary group assignment



Behaviour of method is controlled by two major parameters: (1) total number of decisions trees in the forest and (2) number of variables used for splitting a decision tree at each node (da Silva Júnior et al., 2019; Guan et al., 2013)

Why are these in different groups? They are close together.

Use clusters instead.

Efforts to solve the problem

Researchers from different domains (ecology, neuroscience, meteorology, etc.) have taken different approaches

In meteorology, researchers sometimes select stations where they want to minimize the error (Hutchinson et al., 2009)

Requires expert knowledge of your network and application
If generating surfaces for many days, we can only use stations with unbroken records over that period

In neuroscience (brain imaging), researchers suggested creating larger (randomly assigned) training & testing sets (Little et al., 2017)

Although this method reduces the bias, it will still exist because there are simply more stations in some areas

In ecology, researchers have suggested leaving blocks of samples out (Roberts et al., 2017)

Arbitrary block assignment means that we do not capture the spatial autocorrelation in the error

In ecology, researchers have suggested using a restricted randomization procedure, but this has not been widely applied in meteorology yet (Dale & Fortin, 2002; Fortin & Jacquez, 2000; Roberts et al., 2017)

How do we assign stations to groups in order to implement the procedure?