Monitoring riverine thermal regimes on stream networks: Insights into spatial sampling designs from the Snoqualmie River, WA

Amy Marsha[1], E. Ashley Steel[2], Aimee H. Fullerton[3], and Colin Sowder[1]

[1]Department of Statistics, University of Washington, Seattle WA 98195 USA

[2]Statistics, PNW Research Station, USDA Forest Service, 400 N 34th Street, Suite 201, Seattle, WA 98103 USA

[3] Northwest Fisheries Science Center, NOAA Fisheries Service, 2725 Montlake Blvd. East, Seattle, Washington 98112 USA

[1]Corresponding author: amarsha2@uw.edu

Highlights

- Guidance for monitoring riverine thermal regimes beyond summer mean temperature
- More sites increased predictive precision, but not necessarily predictive accuracy
- Mean temperatures were easier to model than maximums, minimums, or variability
- Winter data were less variable and therefore easier to model than summer data
- Nearby sites with discordant thermal regimes have potential to be highly influential

Abstract:

Understanding, predicting, and managing the spatiotemporal complexity of stream thermal regimes requires monitoring strategies designed specifically to make inference about spatiotemporal variability on the whole stream network. Moreover, monitoring can be tailored to capture particular facets of this complex thermal landscape that may be important indicators for species and life stages of management concern. We applied spatial stream network models (SSNMs) to an empirical dataset of water temperature from the Snoqualmie River watershed, WA, and use results to provide guidance with respect to necessary sample size, location of new sites, and selection of a modeling approach. As expected, increasing the number of monitoring stations improved both predictive precision and the ability to estimate covariates of stream temperature; however, even relatively small numbers of monitoring stations, n=20, did an adequate job when well-distributed and when used to build models with only a few covariates. In general, winter data were easier to model and, across seasons, mean temperatures were easier to model than summer maximums, winter minimums, or variance. Adding new sites was advantageous but we did not observe major differences in model performance for particular new site locations. Adding sites from parts of the river network with thermal regimes which differed from the rest of the network, and which were therefore highly influential, improved nearby predictions but reduced model-estimated precision of predictions in the rest of the network. Lastly, using models which accounted

for the network-based spatial correlation between observations made it much more likely that estimated prediction confidence intervals covered the true parameter; the exact form of the spatial correlation made little difference. By incorporating spatial structure between observations, SSNMs are particularly valuable for accurate estimation of prediction uncertainty at unmeasured locations. Based on our results, we make the following suggestions for designing water temperature monitoring arrays: (1) make use of pilot data when possible; (2) maintain a distribution of monitors across the stream network (i.e., over space and across the full range of covariates); (3) maintain multiple spatial clusters for more accurately estimating correlation of nearby sites; (4) if sites are to be added, prioritize capturing a range of covariates over adding new tributaries; (5) maintain a sensor array in winter; and (6) expect reduced accuracy and precision when predicting metrics other than means.

Key words: water temperature, SSNM, streams, rivers, spatial autocorrelation, monitoring

1. INTRODUCTION

Understanding, predicting, and managing the spatiotemporal complexity of stream thermal regimes on entire stream networks requires carefully designed monitoring strategies. Water temperature regimes on stream networks, influenced by incoming solar radiation, groundwater and atmospheric inputs, as well as a wide range of landscape features such as elevation, human development, riparian vegetation, and geomorphology (Caissie, 2006; Webb et al., 2008), vary within a day and across seasons. These temporal patterns are distributed spatially, with some tributaries experiencing, for example, large daily fluctuations in water temperature during summer and other tributaries experiencing dramatic annual fluctuations (Steel et al., 2016). Capturing the fine-scale temporal variability in temperature at many discrete locations on one stream network is possible using relatively inexpensive in-stream sensors. Site-based measurements can then be used to interpolate particular facets of the thermal regime, e.g., mean summer temperature, to unsampled parts of the network as well as to estimate the effect of variables believed to control water temperature. These models of thermal regimes on stream networks can help identify suitable habitats, prioritize management actions, estimate compliance with legal regulations, and indicate relationships between watershed and instream condition.

As budgets for research, management, and conservation efforts remain limited, new guidance is needed for designing efficient monitoring arrays (a set of spatially distributed monitoring sensors) that capture the spatiotemporal complexity of thermal regimes on the stream network. Moreover, practitioners may wish to understand and predict one or more specific indicators that are of importance for target species and life stages or for protecting thermal regimes through regulatory thresholds. For instance, summer maximum temperatures at least partly determine growth and survival of juvenile salmonids (Satterthwaite et al., 2009) and upriver

migration success for returning adults (Martins et al., 2011). These relatively well-understood physiological relationships have ensured that summer maximum temperature is one of the most commonly evaluated facets of water temperature regimes. However, other facets of the thermal regime may be equally important for species viability. For example, daily fluctuations in winter temperature, when salmonid eggs are incubating in the gravel, are correlated with fry emergence phenology (Steel et al., 2012). Without data on winter variance, ecologists and managers may not be able to account for (or even question) its effect on later life stages. Future monitoring designs may need to be tailored to specifically capture particular facets of the thermal regime and seasons or time windows of interest.

Spatial stream network models (SSNMs) can be fit from water temperature data that were originally collected for other purposes (e.g., Isaak et al., 2011) and not necessarily designed purposefully for building models of water temperature across entire networks. However, ad hoc datasets may not adequately represent spatiotemporal variation in thermal regimes at appropriate scales for managing thermally sensitive species and water uses. Researchers therefore need guidance on necessary sample sizes and best locations for placing additional loggers that will improve predictions and/or estimation of model parameters. Using toy and simulated stream networks, Som et al. (2014) suggest that effective sampling designs should include sites along the full range of important environmental gradients, in major tributaries, in spatial clusters of sites, and at the outlet and headwaters of the stream network. Li et al. (2009) and Zimmerman et al. (2006) found that clustered designs and a mix of space-filling and clustered designs were optimal for similar situations. Falk et al. (2014), using a combination of simulated data on simulated networks and empirical data from the Lake Eacham basin in Queensland, Australia, found that

optimal designs for prediction were distributed fairly evenly over the network but that optimal designs for parameter estimation were somewhat clustered.

In this paper, we use empirical data to expand on the work conducted by Som et al. (2014) and others. We provide practical guidance on the design of monitoring arrays for accurately modeling and predicting particular indicators within complex thermal landscapes. We assess predictive accuracy and estimation of covariate effects from models fit to data from the Snoqualmie River watershed, WA. The models fit in the paper are Gaussian SSNMs, which are geostatistical models that allow for multiple spatially varying random effects (z),

$$Y = X\beta + \sigma_{EUC}z_{EUC} + \sigma_{TD}z_{TD} + \sigma_{TU}z_{TU} + \sigma_{NUG}z_{NUG}$$

where NUG is the nugget effect, and $cor(z_{EUC}) = R_{EUC},\ cor(z_{TD}) = R_{TD},\ cor(z_{TU}) = R_{TU}$ are matrices of autocorrelation values for Euclidean (EUC), TD (tail-down), and TU (tail-up) correlation structures (Peterson and Ver Hoef, 2010). Using these models and our monitoring array of over 40 sensors, we uniquely address the following questions: (I) how big are improvements in model performance with increases in the size of the monitoring array?; (II) where is the best place to add sites to meet particular monitoring goals?; and (III) to what degree does specification of the correlation structure influence the performance of SSNMs? For each question, we explicitly consider whether results differ across facets of the thermal regime (mean, minimum, maximum, and variability) or season (summer and winter). We use empirical data in which the underlying covariance function is not known and, like most rivers, is likely not truly stationary.

2. METHODS

2.1 Study Area

The Snoqualmie River drains a 2,400 km^2 watershed on the west side of the Cascade Range, Washington (Fig. 1). The river begins as three forks whose headwaters lie in mostly forested public land. Just below the convergence of the three forks at the Three Forks Natural Area in Snoqualmie, WA, the river flows over Snoqualmie Falls, a spectacular 82m drop. Below the falls, the river runs through a wide floodplain dominated by agricultural, residential, and commercial land use. Much of this floodplain lies within one of King County's agricultural protection districts. Below the study area, the Snoqualmie River merges with the Snohomish River which drains to Puget Sound shortly thereafter.

2.2 Data

Monitoring sites were located throughout the mainstem and the three main forks of the Snoqualmie River, as well as in the major and minor tributaries (Fig. 1). Practical limitations forced sites to be publicly accessible and within 1 km of a road. The Raging River, a major tributary in the lower watershed, was intentionally oversampled to enable analyses of the effects of scale on monitoring designs in future studies. Thermal regimes on the Snoqualmie River have both a seasonal and daily cycle: though they are fairly messy time series, similar patterns can be observed at a variety of sites on the network (Fig. 2).

For analyses I and III (Table 1), we used empirical data collected every 30-min in summer (May 1, 2014 – August 31 2014) and winter (November 1, 2013 – March 31, 2014). Analysis II relied on data collected every 30-min during shorter but similar time windows in summer (July 1 to August 31, 2014) and winter (January 1 to February 28, 2015) at subsets of the available sites (Table 1). Data going back to July 2011 were available from many of our monitoring sites. At sites

where comparison data were available, data from the same time periods in 2012 and 2013 were visually similar to data used in this analysis and we therefore conclude that this was a typical year.

Data measured within the seasonal windows were summarized by four metrics, each describing a unique facet of the thermal regime. We included a mean (average of all weekly average temperatures; AWAT), a minimum (minimum of all weekly average temperatures; mWAT), a maximum (maximum of all weekly average temperatures; MWAT), and empirical variance (calculated from all observations of the time series; NaiveVar). Prior to calculating summary metrics, data were cleaned to remove missing or erroneous data (Sowder and Steel, 2012). Missing and erroneous data are common with stream temperature data and most often result from loggers coming out of the water during droughts or high flows and recording air temperature.

2.3 Spatial Stream Network Models (SSNMs)

Spatial correlation is the tendency for measurements of the same variable to exhibit similarities as a function of the spatial distance between them. Traditional spatial statistical methods account for the spatial autocorrelation of model residuals via Euclidean distance (straight line distance between locations); however, when working with stream networks this approach may not be ecologically appropriate. For data collected on a river network, spatial stream network models (SSNMs) include more ecologically appropriate covariance structures. These models use moving averages based on stream distance and spatial weights to build statistically valid autocovariance models (Ver Hoef and Peterson, 2010). SSNMs can capture the unique branching structure of the river network, connectivity between sites that are flow-connected, streamflow volume, and directionality of streamflow as well as discontinuities that often occur at river confluences (Cressie et al., 2006; Ver Hoef et al., 2006). The SSNM framework is flexible enough

to allow for a mixture of covariance structures within one statistical model (Peterson and Ver Hoef, 2010).

Models were fit using the SSN package (Ver Hoef et al., 2014) in R statistical software (R Core Team, 2012). In analysis I and II (Table 1), we used an exponential tail-up SSNM. In tail-up SSNMs, the moving average function points in the upstream direction and spatial correlation is restricted to locations that are flow-connected. In analysis III (Table 1), we considered other covariance structures. In all cases, we used mean annual stream flow to determine the spatial weights that split the moving average function at confluences. All models included the same set of covariates: elevation, mean annual flow, and percent commercial land use. Covariates were held constant across models in order to draw conclusions about the effect of logger placement and quantity of loggers used. Models were not intended to be best-fit models, but rather reasonable models that can be used for comparing alternative sampling designs or correlation structures. Models performed similarly with respect to root mean squared error (RMSE), estimated nugget (e.g., remaining unexplained variation) and nugget to sill ratio (e.g., measure of the strength of the spatial dependency), with the exception of winter mWAT and NaiveVar (both seasons) which did not perform as well with respect to these model fit metrics as other season/facet combinations (Table SM1). Parameter estimates, standard errors, and covariance were estimated using restricted maximum likelihood (REML).

2.4 Analysis I: Do we need more sites?

We used a resampling analysis (N=1000 iterations) to quantify the effect, via model estimated coefficients and model predictions, of increasing the number of sites within a monitoring array. For each iteration, a random set of sites was sampled from our existing array and the SSNM

was fit to this subset of the total sites. Resampled arrays included 20, 25, 30, and >30 sites (Table 1). The largest array used three less than the total number of sites available for a particular season after withholding five sites for a prediction analysis. Not all seasonal windows had the same number of sites due to missing data. We compared changes in model performance by metric (AWAT, mWAT, MWAT, and NaiveVar) and by season (summer or winter).

In our resampling approach for analyses I and III, we always included one of the two most downstream sites. First, for ecological reasons, it is difficult to conceptualize a stream network without its most downstream reaches. Second, for practical reasons, these downstream sites are rarely skipped in field-sampling programs. And, third, for statistical reasons, both Som et al. (2014) and Falk et al. (2014) observed that optimal sampling designs include the most downstream monitoring station. To prevent imbalance in the odds of selecting the same monitoring array twice between sampling arrays of different sizes, we identified all possible sets of sampling sites for each array size that also included one of the two most downstream sites. We then selected 1000 possible monitoring arrays, without replacement, from this set of all possible monitoring arrays, ensuring that, for all sizes of monitoring arrays, 1000 unique sets of sampling arrays were selected. We note that because we had a finite number of sites with empirical data, sampling arrays with 20 sites were less likely to contain a similar collection of sites than sampling arrays with 33 sites; however, the effect on the Monte Carlo simulations was small and, to a large degree, reflects the on-the-ground reality of site selection in any particular river network with a finite set of access points.

To explore the effect of adding sites on parameter estimation, we retained the elevation coefficient in each of the above 1000 resampling iterations. We chose the elevation coefficient for exploration because, of the three covariates in our models, elevation has the strongest estimated effect on average water temperature (Steel et al., 2016). We also fit a model with a set of 20 well-

distributed sites and a model with all available sites, each time retaining the modeled coefficient and model-estimated coefficient standard error. To explore the effect of adding sites on predictive accuracy, we identified a set of five sites that were spread across the network, withheld these five sites from all resampling iterations, and compared model predictions for these five sites to empirical observations (Table 1; Fig. 1; Table SM2). Data from the five withheld sites had similar thermal regimes when compared to other sites on the network (Fig. 2). Model predictions at these sites were compared visually and the root mean squared error (RMSE) was calculated to measure the difference between model predictions and the empirical observations by season and array size.

2.5 Analysis II: What is the best location for new sites?

While there may be only a small influence of adding a small number of sites, many on-the-ground practitioners are faced with the question of exactly where to add a few sites when additional funds for monitoring become available. This analysis explores the change in model performance between a model with a base array of sites (n=31 in summer and n=33 in winter) and a model fit with two additional sites. The base monitoring array model was fit using all available sites after removing all of the pairs of sites tested in analysis II. Additionally, we looked at whether the effect of adding particular sites to the monitoring array depended on the metric or season of interest. In this analysis we considered the following four metrics: mean, minimum, maximum, and variance of the empirical data (Table 1). These are similar to AWAT, mWAT, and MWAT but because available data series for this analysis were short, temperature was not summarized weekly before analysis. The data series for this analysis was shorter due to missing or erroneous data at some of the additional sites that were of particular interest, e.g. at tributary confluences and within the spatial cluster.

We considered four approaches for adding two additional sites to a base monitoring array: (1) adding additional sites at tributary confluences (in the tributary and mainstem below the confluence where a site already existed above the confluence), (2) creating a spatial cluster (two sites just upstream of an existing site), (3) adding sites at the spatial extremes of the network (just above furthest downstream site and just below one of the furthest upstream sites), and (4) more densely sampling one tributary with potentially strong influence (adding a tributary and mainstem site to the Raging River) (Fig. 1).

To explore the effect of adding any particular pair of sites on parameter estimation, we retained the model-estimated elevation coefficient and the model-estimated coefficient standard error from each model. To explore the effect of adding any particular pair of sites on predictions, we retained model predictions, model residuals, and model-estimated prediction standard errors for a suite of 12 sites both far away from and nearby each pair of additional sites (Fig. 1). The suite of sites included the most upstream and downstream sites in the monitoring array (far) and sites just above and below each additional pair (near). The residual at a particular site is the difference between the observed value of the metric at that site and the model-estimated value of the metric at that site; positive residuals indicate model under-prediction, whereas negative residuals indicate model over-prediction.

2.6 Analysis III: How do modeling approaches compare?

Practitioners may wonder which covariance structure is best. We used a resampling analysis (N = 1000 iterations) to compare SSNMs with tail-up, tail-down, Euclidean, combined tail-up and tail-down, and combined tail-up, tail-down and Euclidean correlation structures. For completeness, we also compared SSNMs to simple linear models that assume independence. For all models, we compared model performance in terms of parameter estimation and prediction

accuracy, across two sizes of sampling arrays (n=20 and n=33 or 34), four metrics, and two seasons (Table 1).

To explore the effect of model covariance structure on parameter estimation, we retained the elevation coefficient from each model. To explore the effect of modeling approach on predictions, we withheld the same set of five sites as in analysis I and used each model to predict values for these sites by metric and season. We also fit a model with all available sites by metric and season, each time retaining the model prediction and model-estimated prediction standard error at each of the five withheld sites. Model predictive accuracy of each modelling approach was compared by visually estimating whether the bulk of the resampled distribution of predictions covered the true value. Additionally, we assessed predictive accuracy in terms of the distance from the true value of the prediction using all sites, and whether there was appropriate coverage of the model estimated 95% confidence interval.

3. RESULTS

3.1 Analysis I: Do we need more sites?

Parameter estimation. In summer, estimates of the elevation parameter on AWAT increased in precision with increasing sample size (Fig. 3, upper left) as expected. In the absence of a known elevation coefficient, we used the estimated elevation coefficient for an SSNM with all available sites as our best description of the truth (Table SM1). When compared to this best description of the truth, parameters describing the effect of elevation on AWAT were fairly accurate even with smaller sample sizes. In fact, the elevation coefficient from the model using only 20 well-distributed sites was nearly identical to that of the elevation coefficient estimated from all 41 sites; model-estimated standard errors were only somewhat larger for the model built from nearly half

as many sites. However, models built from a random set of 20 sites sometimes estimated coefficients that were fairly far from our best estimate of the truth.

Precision did not increase with sample size at the same rate across all metrics in summer. It was also more difficult for the SSNMs of other metrics to accurately estimate the elevation coefficient at small sample sizes. We refer here not simply to the mean or bulk of the resampled distribution at each array size but to the variability of these resampled estimates, the possibility that an array of a certain size would provide a very in accurate estimate. For example, the effect of elevation on NaiveVar was not well-estimated from a set of 20 well-distributed sites and models built from monitoring arrays that included a random 20 sites were often particularly inaccurate; accuracy did not improve substantially with increasing number of sites (Fig. 3, left).

For a given sample size, there was generally less variability in the estimate of the elevation coefficient in winter than in summer for all metrics except mWAT, where variability was about the same. In winter, estimates of all metrics were fairly accurate (Fig. 3, right). We also observed an increase in precision of the elevation coefficient across monitoring arrays with an increase in sample size; however, the effect was somewhat less dramatic in winter than in summer, perhaps because coefficients were relatively more precise at smaller sample sizes. SSNMs of mWAT showed the most dramatic increase in precision with increasing sample size (Fig. 3, right).

Prediction. In summer, precision of predicted AWAT at unmeasured sites increased with sample size (Fig. 4, upper left). For some sites, e.g., NF County Bridge and Raging Bridge, accuracy also seemed to increase with increasing sample size but for other sites, e.g. Tokul and Taylor, accuracy seemed to decrease. Although AWAT was not perfectly predicted at any of our five test sites, it was reasonably well-predicted for all of them. Looking across metrics in summer, mWAT was

adequately predicted but MWAT and NaiveVar were not well-predicted at most of the five sites, and particularly poorly predicted at some sites. Looking at summer metrics, the root mean squared error (RMSE) averaged across sites was largest when modeling MWAT and NaiveVar (Table SM2) regardless of sample size. The SSNMs tended to under-predict summer MWAT at all five sites. And metrics describing the same site were not systematically under- or over-predicted. For example, AWAT on Taylor River was over-predicted but MWAT was under-predicted (Fig. 4, left).

In winter, predictions of AWAT were also reasonably accurate at small sample sizes and increased in precision with increasing sample size. For AWAT in winter, RMSE averaged across all five sites was about the same when increasing from 20 sites to 34 sites; 0.306 and 0.246 respectively (Table SM2). Looking across metrics in winter, accuracy did not necessarily increase with increased sample size. As sample size increased, predictive accuracy of NaiveVar at Raging Bridge became noticeably worse. When comparing one metric in winter to the same metric in summer, there was generally greater predictive accuracy in winter for all sites (Fig. 4; Table SM2). Also comparing across seasons, there were shifts in under- versus over-prediction for a given metric at a given site (Fig. 4).

3.2 Analysis II: What is the best location for new sites?

Parameter estimation. In both July through August and January through February, estimates of the elevation parameter on mean temperature were similar regardless of which two new sites were added to the monitoring array. The same was observed for other metrics, regardless of season (Fig. SM3).

Prediction. There was very little difference in predicted mean summer temperature (or model residuals) for any site, when two new sites were added to the monitoring array. This result was consistent whether sites were added at tributary confluences, in a cluster, at the tips of the network, or within a particular subbasin (Fig. 5, top and middle). Tiny shifts were detectable in predictions for some sites when two new sites were added. For example, the predicted value for the most upstream site on the Tolt (T1) was a bit warmer when two sites surrounding where the Tolt River joins the mainstem Snoqualmie River were added to the monitoring array (Fig. 5, top and middle).

The standard errors of model predictions did change as sites were added to the monitoring array (Fig. 5, bottom). Adding additional sites in the Raging River (Add Dense in Fig. 5) resulted in a greater range of prediction standard errors across our set of example sites. Interestingly, when two sites surrounding the Tolt River confluence with the mainstem were added to the monitoring array, there was a marked increase in prediction standard error across all sites, even sites far from the Tolt River confluence (Fig. 5, bottom). In exploring this result, we found that the thermal regimes of the three sites surrounding the confluence are very different from one another (Fig. 1) and, furthermore, that when we added of a pair of close-in-space yet similar sites along with the Tolt River confluence sites, this increase in prediction standard error was ameliorated (results not shown).

Comparing model performance across metrics by looking at model residuals (Fig 5, middle; Fig. 6), there were no major differences when two sites were added, regardless of which sites were added, which metrics were being considered, or which site was estimated. As observed for July through August mean temperature, adding sites from the Tolt River confluence increased some but not all of the July through August maximum temperature predictions (Fig. 6, middle).

In winter, most patterns were similar, except in the Raging River, where the effect was in the opposite direction. A denser set of sensors in the Raging River resulted in higher winter mean and winter maximum temperature predictions at most sites throughout the watershed (Add Dense in Fig. SM1; Fig. SM2). The inclusion of the spatial cluster sites, sites at the Sunday River confluence, or sites at the far ends of the network (e.g., UpDown sites) noticeably reduced mean temperature prediction standard errors across the network, though the predicted mean temperatures themselves were not very different from those of the model with the base array of sites (Fig. SM1).

3.3 Analysis III: How do modeling approaches compare?

Parameter estimation. In summer, estimates of the elevation parameter on AWAT were similar across modeling approaches, using either 20 or 34 sites, and across the four model structures (Fig. 7, top left). When comparing across metrics in summer, the linear model estimated a stronger negative relationship between elevation and mWAT than did any of the SSNM estimates, and this effect persisted across the two sample sizes. The tail-up, tail-down, and Euclidean mixed-model estimated an elevation parameter on summer NaiveVar very similar to the other models, however, when the number of sites was increased to 34 this model estimated a notably weaker positive relationship between elevation and summer NaiveVar than the other models (Fig. 7, left).

In winter, models disagreed slightly in estimates of elevation's effect on AWAT and MWAT, whereas they agreed very well for mWAT and NaiveVar (Fig. 7, right). When modeling winter AWAT and winter MWAT, the linear model estimated a stronger negative relationship with elevation than the SSNMs.

Prediction. In summer, the tail-up SSNM tended to result in a more accurate AWAT prediction at four of the five sites than did the linear model using either 20 or 34 sites (Fig. 8, top left). The only

strong pattern was, as expected, that the coverage of the model-estimated confidence interval from the linear model was generally much poorer than that of any of the spatial models (Fig. 9). However, there was no one covariance structure that always resulted in a more accurate AWAT prediction (Fig. 9, top left) and the predictive accuracy varied quite a bit between the different SSNMs. When comparing across metrics and sites in summer, there were no consistent patterns. For example, the linear model had better predictive accuracy for Tokul Creek than the SSNMs when modeling summer mWAT, while the SSNMs had better predictive accuracy when modeling summer MWAT (Fig. 9, left). At the smaller sample size, the tail-up SSNM often had more variability in metric predictions at the five sites than the linear model; particularly when modeling MWAT and NaiveVar (Fig. 8, left).

In winter, again, predictions from the SSNMs were not uniformly more accurate than those from the linear models but the coverage of the model-estimated confidence interval was much better for SSNMs (Fig. 9, right). There was little difference across correlation structures (Fig. 9, right).

4. DISCUSSION

Arrays of temperature sensors for measuring and modeling stream temperature are being installed on many river networks. Our resampling study provides guidance with respect to sample size, locating new sites, and selection of a modeling approach across multiple facets of the thermal regime and two seasons. This guidance is intended for networks on which data do not yet exist. We have demonstrated that, as expected, increasing the number of monitoring stations improves both predictive precision and the ability to estimate covariates of stream temperature; however, even relatively small numbers of monitoring stations, n=20, can do an adequate job when well-

distributed and when used to build models with only a few covariates. However, particular caution is necessary with small arrays. For example, for some arrays with n=20, the predicted metric value was quite off relative to larger sample sizes (Fig. 4). In general, winter indicators on the Snoqualmie River are easier to model than summer indicators and mean temperatures in both seasons are easier to model than maximums, minimums, or variance. Adding new sites is advantageous but we did not observe major differences in model performance as a result of exactly where new sites were added, except that adding sites which are close together in network space but which differ in their thermal regime reduces model-estimated predictive accuracy across the network. Lastly, using models which account for the network-based spatial correlation between observations made it much more likely that estimated prediction confidence intervals covered the true parameter, but the exact form of the spatial correlation made little difference. We note that these findings are particular to the Snoqualmie River; they are likely to be useful in other river systems with similar thermal regimes and highly spatially-structured covariates. In the future, results from similar studies across a range of river networks can provide more global guidance.

4.1 Implications of variation in monitoring array performance across sample size, metric, and season

Sample Size. As expected, precision of the elevation parameter and precision of predictions at unsampled sites improved with increasing sample size. However, even monitoring arrays of only 20 sites were relatively unbiased with respect to parameter estimation, and most sets of 20 sites performed similarly to the full set of 41 sites for summer mean temperature. Because we used empirical data, we did not have access to the true effect of elevation for an estimate of the accuracy of the model estimation procedure; yet, we did not observe a shift in the parameter estimate with increasing sample size for any metric or season. Such a shift would be an indication of a change

in the magnitude of the parameter estimate and therefore a change in accuracy. Accuracy of parameter estimates and of predictions at unsampled sites rarely improved with increased sample sizes, but there were wide differences across facets of the thermal regime, season of monitoring, and site being estimated. For example, when considering summer AWAT, the RMSE across five unsampled sites decreased by 19.8% when increasing the sample size from 20 to 34 sites, while winter NaiveVar decreased by 32.5% (Table SM2).

Interestingly, Sály and Erös (2016), in an investigation of the effect of sample size and sampling design on ordination-based variance partitioning of data collected on a dendritic network, found that increasing sample size did little to reduce residual error. In their analysis, the primary effect of increasing sample size was to decrease the variance explained purely by environmental covariates and increase the variance explained by the interaction of environmental covariates and spatial structure. Though they caution that such results may be specific to their study site, such results are likely quite universal on river systems because of the strong underlying spatial correlations between drivers of instream condition and the river network itself (Lucero et al., 2011). It is difficult, in fact, to think of any environmental covariate, e.g. geology, elevation, mean annual stream flow, percent agriculture, road density, which might be randomly distributed across a river network aside from, in very extreme situations, point-source anthropogenic pollution. While the structure of water temperature on a river is specific to the system being investigated, the spatial correlation in the data will necessarily increase with increasing sample size, assuming sites are reasonably distributed across available space. Take as an extreme example, data from three sites distributed across a river network. It is quite possible for these observations to be relatively uncorrelated. Now consider data from 100 sites distributed on the same river network; unless the river network is gigantic, it is highly unlikely that these data do not exhibit spatial structure.

Although we did not test this specifically, sample size, or rather sample density, is likely to affect estimates of the degree of spatial correlation between observations on river systems (Sály and Erös, 2016) as it does in other spatial contexts (Zhu and Zhang 2006).

Metric. While means were relatively easy to understand and predict, other facets of the thermal regime were more complicated. For some sites, predictions of one facet were relatively accurate while predictions of other facets were not. Overall, summer maxima tended to be under-predicted and estimates of summer variance were inaccurate at all sites tested. There has been an increasing emphasis on measuring, monitoring, and understanding patterns in thermal variability (Arismendi et al., 2013). Understanding covariate effects on variability and predicting variability at unsampled locations may be more difficult than similar analyses on mean temperatures.

Model performance for any particular facet of a particular stream network will be a function of the distribution of the facet, the strength of the relationship between available covariates and the facet, the spatial variability of that facet, and the spatial variability of the thermal regime in that river system. In our analysis, we considered a set of three covariates with widely understood relationships to thermal regimes; however much of the work has been done on summer mean temperature and these same covariates are not necessarily as strongly related to other facets of the thermal regime. Where covariates are poorly correlated with the response of interest, modeled predictions for new areas will all tend toward the overall mean. As more research is completed on landscape factors that influence minimums, maximums, and variability in thermal regimes, modeled predictions are likely to become more accurate overall. Statistical developments will also contribute to improved models. The Gaussian SSNMs fit in this paper assume normally distributed residuals, which is likely not the case when modeling extremes. So when using SSNMs to model

extremes or temperature variability, exercise caution and take steps to evaluate model performance.

Season. Historically, most stream temperature measurements have been recorded in summer; however, riverine thermal regimes on the Snoqualmie River and on similar temperate rivers are likely easier to model in winter because most facets of the thermal regime show less spatial and temporal variability in winter than in summer (Steel et al., 2016). According to the high estimated nugget to sill ratio for 3 of the 4 winter metrics, the tail-up SSNMs we fit indicate a weaker spatial dependency than the corresponding metrics in summer (Table SM1). Predictions of all four facets, even minimum temperature, at all five withheld sites, were more accurate in winter than in summer (Table SM2). These results are extremely helpful because often there are fewer loggers in winter monitoring arrays. Keeping temperature loggers installed successfully in winter is more challenging than in summer as snow may prevent access to sites for checkup visits and high flows from winter storms can wash loggers out of the water or even wash the logger and the entire anchoring system, tree or bolder, downstream. Monitoring programs interested in minimum temperatures, however, generally do need to have a large number of loggers recording during winter months. Precision of winter minimum temperature predictions at most of the five unsampled sites increased significantly with increasing number of sites; indeed, prediction of winter minimum temperature varies greatly for most sites at the smaller monitoring array sizes. While larger sample sizes are needed for monitoring programs designed to capture minimum temperatures, there appears to be only a weak effect of elevation on minimum temperature in winter, making access to sites higher in the watershed not quite as essential. Although the weak effect of elevation in winter has only been documented for this river basin, similar results are likely

for other networks; reductions in thermal variance at cold temperatures are driven by the inability of flowing waters to drop far below zero and the buffering effects of snow fall and snow melt.

The effects of minimum temperature have been less well-studied for most aquatic species than the effects of maximum temperature, but evidence is mounting that for some species and in some areas, minimum temperatures can be limiting (Jonsson and Jonsson, 2009) and we expect climate change to have a pronounced effect on minimum temperatures in this region (Arismendi et al., 2013). Better estimates of minimum temperatures across stream networks and an understanding of correlates and drivers of minimum temperatures may require an increased density of temperature loggers in many monitoring arrays. In particular, having more loggers at spatial locations spread across a greater range of available covariates.

Site. On the Snoqualmie River, as is common on other rivers, some sites were simply different from the rest of the river network and these differences made them difficult to model. We found for example that Lower Cherry was unusual and more difficult to predict across all four facets of the thermal regime in summer and for both maximum and minimum winter temperatures. The full network includes a site not too far upstream on Cherry Creek, but there are changes in land-use and land-form, i.e. increases in small farms and decreases in hillslope, as the creek moves downstream and there may be inputs of colder or warmer water (e.g., subsurface seeps) that make facets of the summer thermal regime at the lower site difficult to predict from nearby data. Facets of the thermal regime that are patchy and that are not well-correlated with commonly-used two-dimensional correlates, e.g., elevation or land-use, will simply be challenging to model and to predict at unsampled sites.

Biologists and statisticians building and using predictions from SSNMs sometimes have a latent belief that the models are evenly inaccurate over space and time. This is, of course, not true. We have demonstrated that the model tends to over or under predict some metrics and some sites, no matter where on the stream network the data are collected. While the details may vary from site to site and from network to network, the more general result that model accuracy varies on the network likely holds in all basins. Looking again at Cherry Creek where the thermal regime is warmer most of the year than even the Lowest Mainstem, highly variable in summer, and less variable in winter (Fig. 2), we found that the SSNM always predicted summer water temperatures that were much cooler than what was observed (Fig. 4). Unless a covariate is included that can explain the warm summer temperatures at Cherry Creek, the model will fall short.

Winter temperatures on Lower Cherry were more similar to those observed in other parts of the network and the model, not surprisingly, did much better at predicting averages and variance. Unusual sites that are cooler than expected, for example the aptly named Icy Creek which drains into the warm Raging River, or Taylor River which is cooler than might be expected from its location in the network (Fig. 2) will also always be poorly predicted unless the underlying processes driving the cool temperatures are captured by the covariates in the model.

4.2 The effect of adding particular sites to the Snoqualmie River monitoring network

While the addition of new sites is clearly advantageous for any monitoring array, there were few observed differences in model performance based on which particular sets of sites were added. Model-estimated elevation coefficients for any of the four facets and model residuals for any of the four facets and for any of the twelve example sites changed very little no matter which

two sites were added to the array, suggesting that adding easy access sites, wherever they are, should be considered. However, a few surprising and helpful insights did emerge.

First, we note that additional sites on the Raging River (Add Dense, Fig. 6, middle) decreased many predictions of maximum temperature in other parts of the network, even at sites fairly far from the Raging River. The two added sites included an additional warm site on the mainstem Raging, very close to similar sites, and an unusually cold small tributary to the Raging. It is intuitive to want to include sites in parts of a river network that are unusual in some way, for example, cold and stable ground-water fed sites. However, adding these unusual sites may shift predictions of particular facets at sites across the river network due to both model covariates and spatial structure. The trade-off of including such sites will often be valuable, but it will be important for managers and modelers to explore these possibilities when selecting sites and interpreting model results.

Second, the addition of pairs of sites that are nearby in space yet which have very different thermal regimes from each other, perhaps as a result of a cold water tributary or a point source input of warmer water, may increase the prediction standard errors across the entire network. In our analysis, the addition of just two sites around the Tolt River confluence, where it joins the mainstem Snoqualmie River (Fig. 1), radically increased model estimated prediction standard errors even at the lowest mainstem site, the furthest upstream tributaries, and at flow unconnected sites. This result likely arises because the addition of the Tolt River tributary sites forced a fairly large reduction in the estimated spatial covariance of nearby sites and therefore reduced model confidence across the network. This effect was slightly diminished with the addition of a set of two nearby clustered sites that were, in fact, very similar to one another and to a third nearby site on the Raging River. The addition of this second spatial cluster increased the modeled estimate of

spatial covariance at nearby sites, reducing the effect of the Tolt River tributary. A similar increase in prediction standard errors across the network when adding the Tolt tributary sites was also observed when considering the other three metrics (max, min, and variance), though it was most pronounced when modeling mean and maximum temperatures. The addition of new sites to an existing array may therefore impact models and estimates of some facets of the thermal regime differently than models and estimates of other facets.

Adding sets of sites that are close together in space (spatial clusters) is important for a clear understanding of spatial heterogeneity in thermal regimes and for estimating the left side of the semivariogram (Som et al., 2014). Design of monitoring arrays also needs to consider that, especially when there is only one or just a few such clusters of sites, these clusters will strongly influence estimates of spatial structure across the river network. While application of SSNMs assumes stationarity of the correlation structure, there could be few natural rivers for which this assumption holds perfectly. Consider two sites located 100 m apart on almost any mainstem; all facets of water temperature regimes are relatively highly correlated between these sites even after covariates are incorporated into the model. Now consider any place on the same river network where a cool tributary flows into a warmer tributary and imagine two sites that are also 100 m apart but with one site situated on each of the two tributaries. There will be much less correlation in thermal means or in any other facet of the thermal regime between these two sites even after covariates are incorporated. While the performance of these models is strong even when assumptions are not met perfectly, it is important to evaluate the application of these models using empirical data from natural systems. The improved estimation of the spatial heterogeneity as a result of inclusion of spatial clusters in the monitoring design will be reflected in a shift of the prediction standard errors across the network, and not just the sites near the spatial cluster.

Maintaining these spatial clusters long-term is a challenge due to the higher probability of at least one sensor in a spatial cluster failing (a minimum of 3 are needed) and therefore an important consideration in monitoring array design and maintenance.

4.3 Guidelines for selecting a modeling approach

Application of SSNMs to summer mean temperatures, thereby accounting for the spatial correlation in the data, has been shown to significantly improve accuracy of predictions of summer mean and maximum temperatures at unsampled locations (Isaak et al., 2010; Ruesch et al., 2012). However, in selecting a modeling approach, there are two decisions that need to be made. The first is whether to fit a standard linear model which assumes independent errors or to fit a SSNM that addresses the spatial autocorrelation between sites on a branching river network. If the SSNM approach is chosen, one must then decide what type of spatial covariance structure to impose on the stream network.

In our analysis, the predictions at unsampled sites from the SSNMs did not uniformly have greater accuracy than the predictions from the linear models. However, the prediction intervals around the SSNM predictions typically covered the true parameter, whereas the prediction intervals from the linear model did not. Due to the spatial dependence in the data, the linear model assumes there are more independent samples, and thus a bigger effective sample size then there truly is. As a result, the resultant predictions from the linear models tend to be over-confident; they do not have appropriate coverage of the true parameter.

One possible explanation for why the SSNMs did not uniformly have greater accuracy than linear models is that different facets of the thermal regime have more and less spatial covariance. Variance decomposition is used to attribute the total amount of variation in the model response

variable to particular sources, including model covariates and residuals. Decompositions using SSNMs include those sources, but also include the spatial network structure. By decomposing the variance, we gain insight into what processes are at work in our network and into the relative strength of each of those processes. Using the data from our river network, Isaak et al. (2014) used wavelet metrics on intra and inter-daily time steps to fit both nonspatial models and SSNMs, and decomposed the variance in order to analyze the relative strength of the spatial structure. The results of that analysis suggested that stream temperature fluctuations on the Snoqualmie River have a strong spatial component over short periods (intra-daily), but a weak spatial component over longer periods (inter-daily) (Isaak et al., 2014, Figure 6). The metrics used in our analysis were on a weekly time step (i.e. average weekly average temperature (AWAT)), and according to the variance decomposition, after accounting for the spatially structured covariates, the relative strength of the network spatial structure was not as strong. This could explain why SSNMs, which accounted for spatial network structure, did not uniformly perform better.

The SSNMs tended to have greater predictive accuracy than the linear model when modeling summer MWAT, which, of the summer metrics included in our analysis, has one of the strongest spatial structures (Steel et al., 2016) and also has clearly understood biological (Ebersole et al., 2001) and regulatory implications (Poole et al., 2004; Ruesch et al., 2012). So if a manager is interested in thermal maxima, using a model that incorporates the spatial network structure is particularly important in getting a prediction with good coverage. Specific results may differ in other basins, particularly where the best set of covariates contains little spatial structure. For other applications, an assessment of the estimated covariance parameter can help guide the choice of a spatial versus non-spatial model.

Once the decision was made to fit a SSNM, we observed very few differences in elevation parameter estimates or metric predictions at unsampled sites between the alternative exponential covariance structures. Even at the small network size of 20 sites, the elevation parameter estimation was approximately the same regardless of whether we used the exponential tail-up, exponential tail-down, or the mixture exponential tail-up, tail-down, and Euclidean models. Given that no major differences were observed between the SSNMs, there appears to be no penalty in adding the additional covariance components. Covariance doesn't suffer from the problem of overfitting that happens when including additional covariates. It is also worthy of note that when fitting SSNMs with a mixed covariance structure, there is no need to determine a priori which covariance models to include (Frieden et al., 2014).

4.4 Guidelines for selection of water temperature monitoring sites on a river network

Combining results from Som et al. (2014), who used a wide variety of toy and simulated river networks, with our results, based on empirical data from one year and one network, a few key principles can be distilled. First, pilot data distributed across the network is extremely useful. Such pilot data can help identify river reaches with thermal regimes that differ from the rest of the network. These reaches have the potential for high leverage and are recommended as good sample sites (Som et al., 2014). Pilot data might be available from historical records, a small set of loggers deployed across the network in advance of the design of the full monitoring array, remotely-sensed data, or even spot data collected through quick visits to remote parts of the network on a few days with similar cloud cover and precipitation histories.

Second, maintaining a distribution of sample sites across the entire network, from upstream to downstream, across the full range of covariates, and in a balance of areas with high leverage is

ideal. In terms of long-term maintenance, access issues often preclude high elevation or remote sites in the winter months. Installing loggers at these challenging sites anyway and accepting some logger loss in winter may be worth the trade-off for improved precision of estimates in summer, and in particular, if winter minimums are of interest. Areas with high leverage are generally those that are much cooler or much warmer than other sites with similar network position and covariates; they may be difficult to identify in advance. Ideally, new covariates can eventually be identified that explain unusual patterns in water temperature regimes.

Third, one cluster at the top and bottom is recommended by Som et al. 2014, but we note that nearby sites which differ from one another can have a very strong influence with respect to estimating spatial correlation. A distributed set of clusters from outlet to headwaters may be ideal, even though unintuitive. Managers may wonder, why "waste" loggers measuring temperatures that we know are similar to those measured by a nearby logger; yet, those clusters provide important information to the model about just how similar nearby sites may be with respect to particular indicators and take little extra effort to maintain because they can all be accessed during one site visit. Only one or two such clusters, located accidentally in areas of high or low spatial covariance could be dangerous. A larger set of clusters reduces the risk of incorrectly applying a particularly low or high estimate of spatial covariance across the network. Another good argument for maintaining several clusters is that the logistics of keeping all three loggers that form a cluster in the water are challenging. The loss rate for a cluster of three sites is three times higher than the expected loss rate for just a single logger. So, maintaining a monitoring program with a few spare clusters for challenging years is a good idea.

Fourth, adding more sites is always a good idea even if they are not in ideal locations. We found no evidence that, for the Snoqualmie River, it mattered much which sites were added. We

had just over 40 temperature loggers dispersed throughout the river network. According to Isaak et al. 2014, a minimum of 50 loggers are needed to fit SSNMs with multiple covariates. We didn't have access to more empirical data, however many people are making decisions based on approximately 20-40 loggers. The number of loggers that need to be deployed in a network will depend on the spatial heterogeneity in that network, the strength of the covariates, and the size of the network. Larger networks will be better modeled with more loggers.

Capturing tributary confluences, while intuitively important, may not be essential. In Som et al. (2014), confluence-focused clusters tended to be the optimal design in inference regarding estimation of an overall mean for tail-up spatial processes. Although we did not estimate an overall mean in our analyses, we did not find a strong relative importance of including these confluence triads in terms of model performance. Rather, when adding more sites, it may be best to spread these sites out across a greater range of model covariates which are likely to influence the facet of interest (Jackson et al., 2016).

Lastly, more sites will be required to estimate some facets than others. If extremes or variability are indicators of important biological phenomena, more sites will be needed than for means. Advances in our understanding of the ecological drivers of these facets and modeling advances for describing unusually distributed facets on river networks will improve our ability to monitor the full spatiotemporal complexity of riverine thermal regimes.

Acknowledgements

This work was supported by the USDA Forest Service, PNW Research Station.

REFERENCES

Arismendi, I., Johnson, S.L., Dunham, J.B., Haggerty R., 2013. Descriptors of natural thermal regimes in streams and their responsiveness to change in the Pacific Northwest of North America. Freshw. Biol. 58 (5), 880-894.

Caissie, D., 2006. The thermal regime of rivers: A review. Freshw. Biol. 51 (8), 1389-1406.

Cressie, N., Frey, J., Harch, B., Smith, M., 2006. Spatial prediction on a river network. J. Agric. Biol. Environ. Stat. 11 (2), 127-150.

Ebersole, J.L., Liss, W.J., Frissell, C.A., 2001. Relationship between stream temperature, thermal refugia and rainbow trout Oncorhynchus mykiss abundance in arid-land streams in the northwestern United States. Ecol. Freshw. Fish. 10 (1), 1–10.

Falk, M.G., McGree, J.M., Pettitt, A.N., 2014. Sampling designs on stream networks using the pseudo-Bayesian approach. Environ. Ecol. Stat. 21 (4), 751-773.

Frieden, J.C., Peterson, E.E., Webb, J.A., Negus, P.M., 2014. Improving the predictive power of spatial statistical models of stream macroinvertebrates using weighted autocovariance functions. Environ. Modell. Softw. 60, 320-330.

Isaak, D.J., Luce, C.H., Rieman, B.E., Nagel, D.E., Peterson, E.E., Horan, D.L., … & Chandler, G.L., 2010. Effects of climate change and wildfire on stream temperatures and salmonid thermal habitat in a mountain river network. Ecol. Appl. 20 (5), 1350-1371.

Isaak, D.J., Wenger, S.J., Peterson, E.E., Ver Hoef, J.M., Hostetler, S., Luce, C.H., … & Wollrab, S., 2011. NorWeST: An interagency stream temperature database and model for the Norwest United States. US. Fish and Wildlife Service, Great Northern Landscape Conservation Cooperative Grant. Project website: www.fs.fed.us/rm/boise/AWAE/projects/NorWeST.html.

Isaak, D.J., Peterson, E.E., Ver Hoef, J.M., Wenger, S.J., Falke, J.A., Torgersen, C.E., … & Ruesch, A.S., 2014. Applications of spatial statistical network models to stream data. Wiley Interdisciplinary Reviews: Water. 1 (3), 277-294.

Jackson, F.L., Malcolm, I.A., Hannah, D.M., 2016. A novel approach for designing large-scale river temperature monitoring networks. Hydro. Res. 47 (3), 569-590.

Jonsson, B., Jonsson, N., 2009. A review of the likely effects of climate change on anadromous Atlantic salmon Salmo salar and brown trout Salmo trutta, with particular reference to water temperature and flow. J. Fish Biol. 75 (10), 2381-2447.

Li, J., 2009. Spatial multivariate design in the plane and on stream networks. PhD Thesis, University of Iowa.

Lucero, Y., Steel, E.A., Burnett, K.M., Christiansen, K., 2011. Untangling human development and natural gradients: Implications of underlying correlation structure for linking landscapes and riverine ecosystems. River Systems. 19 (3), 207-224.

Martins, E.G., Hinch, S.G., Patterson, D.A., Hague, M.J., Cooke, S.J., Miller, K.M., … & Farrell, A.P., 2011. Effects of river temperature and climate warming on stock-specific survival of adult migrating Fraser River sockeye salmon (Oncorhynchus nerka). Glob. Change Biol. 17 (1), 99-114.

Peterson, E.E., Ver Hoef, J.M., 2010. A mixed-model moving average approach to geostatistical modeling in stream networks. Ecology. 91 (3), 644-651.

Poole, G.C., Dunham, J.B., Keenan, D.M., Sauter, S.T., McCullough, D.A., Mebane, C., … & Deppman, M., 2004. The case for regime-based water quality standards. BioScience. 54 (2), 155–161.

R Core Team, 2012. R: A language environment for statistical computing. R foundation for statistical computing, Vienna, Austria. http://www.R-project.org/, accessed August 2016.

Ruesch, A.S., Torgersen, C.E., Lawler, J.J., Olden, J.D., Peterson, E.E., Volk, C.J., & Lawrence, D.J., 2012. Projected Climate-Induced Habitat Loss for Salmonids in the John Day River Network, Oregon, USA. Conserv. Biol. 26 (5), 873-882.

Sály, P., Erös, T., 2016. Effect of field sampling design on variation partitioning in a dendritic stream network. Ecol. Complex. 28, 187-199.

Satterthwaite, W.H., Beakes, M.P., Collins, E.M., Swank, D.R., Merz, J.E., Titus, R.G., Sogard, S.M., Mangel, M., 2009. Steelhead life history on California's central coast: insights from a state-dependent model. T. Am. Fish. Soc. 138 (3), 532-548.

Som, N.A., Monestiez, P., Ver Hoef, J.M., Zimmerman, D.L., Peterson, E.E., 2014. Spatial sampling on streams: principles for inference on aquatic networks. Environmetrics. 25 (5), 306-323.

Sowder, C., Steel, E.A., 2012. A note on the collection and cleaning of water temperature data. Water. 4 (3), 597-606. Available at http://www.mdpi.com/2073-4441/4/3/597

Steel, E.A., Tillotson, A., Larsen, D.A., Fullerton, A.H., Denton, K.P., Beckman, B.R., 2012. Beyond the mean: the role of variability in predicting ecological effects of stream temperature on salmon. Ecosphere. 3 (11), 1-11.

Steel, E.A., Sowder, C., Peterson, E.E., 2016. Spatial and temporal variation of water temperature regimes on the Snoqualmie River network. J. Am. Water Resour. Assoc. 52 (3), 769-787.

Ver Hoef, J.M., Peterson, E., Theobald, D., 2006. Spatial statistical models that use flow and stream distance. Environ. Ecol. Stat. 13 (4), 449-464.

Ver Hoef, J.M., Peterson, E.E., 2010. A moving average approach for spatial statistical models of stream networks. J. Am. Stat. Assoc. 105 (489), 6-18.

Ver Hoef, J.M., Peterson, E.E., Clifford, D., Shah, R., 2014. SSN: An R package for spatial statistical modeling on stream networks. J. Stat. Softw. 56, 1-43.

Webb, B.W., Hannah, D.M., Moore, R.D., Brown, L.E., Nobilis, F., 2008. Recent advances in stream and river temperature research. Hydrol. Process. 22 (7), 902-918.

Zhu, Z., Zhang, H., 2006. Spatial sampling design under the infill asymptotic framework. Environmetrics. 17 (4), 323-337.

Zimmerman, D.L., 2006. Optimal network design for spatial prediction, covariance parameter estimation, and empirical prediction. Environmetrics. 17 (6), 635-652.

Figure Legends

Figure 1: Map of Snoqualmie River, Washington, USA. Sites withheld to test predictive accuracy in resampling analyses, evaluating effect of sample size, and comparing modeling approaches (Analysis I and III; Table 1; Figure 2) are identified with an inner dot. Sites systematically added to explore how the addition of particular sets of sites affects model performance (Analysis II; Table 1) are identified with solid symbols: star, triangles, pentagons, circles. Sites in which model performance was evaluated in Analysis II are labeled with a short site name which is also used in Figures 5 and 6. Time series of temperature data for five sites associated with the confluence of the Tolt and Snoqualmie Rivers are inset to display differences in data for nearby sites. The two sites added in Analysis II (Table 1, Figures 5 and 6) for the Tolt River confluence are identified as solid triangles in inset which, unlike the triangles in the main figure, represent just one site each.

Figure 2: Observed data for the five sites withheld to test predictive accuracy in resampling analyses, evaluating effect of sample size, and comparing modeling approaches (Analysis I and III; Table 1). These five sites are spatially identified in Figure 1. Only data collected at 6am and 6pm are displayed for clarity. Observed data from the furthest downstream site and from the furthest upstream sites on the North Fork, South Fork, Middle Fork, Tolt River, and Raging River are displayed in grey for context.

Figure 3: Elevation parameter estimates from SSNM resampling analysis (Analysis I) varying sample size, season (summer and winter), and metric (AWAT, mWAT, MWAT, NaiveVar). The bars for N=20 and N=41 or 42 in each panel show the estimated elevation parameter using either a well-distributed set of 20 sites or all available sites, along with the 95% confidence interval for

each estimate. The dashed line corresponds to the estimate from the model using all available sites and is included for easy comparison across sample size.

Figure 4: The model-predicted metric values at the five left-out sites from the SSNM resampling analysis (Analysis I). Predictions at each site are from models that vary by sample size, season (summer and winter), and metric (AWAT, mWAT, MWAT, NaiveVar). Predictions are compared to the observed value at each site (solid line).

Figure 5: The July – August mean SSNM predictions, residuals, and prediction standard errors from each 'add two sites' model and from a base monitoring array model which did not include any of these additional sites. Model predictions, residuals, and prediction standard errors are reported at a suite of 12 sites which include the most upstream and downstream sites, a flow unconnected site (far), and sites above and below the added pairs (near). The nearby sites corresponding to each added pair of sites are indicated by filled in circles. All sites are labeled corresponding to their location in the network (i.e. middle fork, mainstem, or tributary name) and to their position in the direction of water flow (high numbers being more downstream and low numbers being more upstream). Added sites and predicted sites are further identified in Figure 1.

Figure 6: The July – August minimum, maximum, and variance SSNM residuals from each 'add two sites' model and from a base monitoring array model which did not include any of these additional sites. Model residuals are reported at a suite of 12 sites which include the most upstream and downstream sites, a flow unconnected site (far), and sites above and below the added pairs (near). The nearby sites corresponding to each added pair of sites are indicated by filled in circles.

All sites are labeled corresponding to their location in the network (i.e. middle fork, mainstem, or tributary name) and to their position in the direction of water flow (high numbers being more downstream and low numbers being more upstream). These sites are further identified in Figure 1. The gray horizontal lines correspond to the zero residual line, indicating a perfect prediction.

Figure 7: Elevation parameter estimates from linear models and SSNMs using randomly selected sites for two sample sizes. Models varied by season (summer and winter), metric (AWAT, mWAT, MWAT, NaiveVar), and correlation structure. U is tail-up correlation, D is tail-down correlation, UDE is combined tail-up, tail-down, and Euclidean correlation, and I is the linear model with an independent correlation structure (no spatial correlation). Random sampling was done at sample sizes of 20, and three sites less than the total number of available sites after removing the five withheld sites (N=33 in summer; 34 in winter).

Figure 8: The model predicted metric values at five withheld sites from linear models (I) and exponential tail-up SSNMs (U) where the sites included in the model were randomly selected at two sample sizes. Sites and models are compared during two seasons (summer and winter), and for four metrics (AWAT, mWAT, MWAT, NaiveVar). Random sampling was done at monitoring array sizes of 20 and 33 in summer, and 20 and 34 in winter. The solid horizontal lines represent the observed metric value at that site.

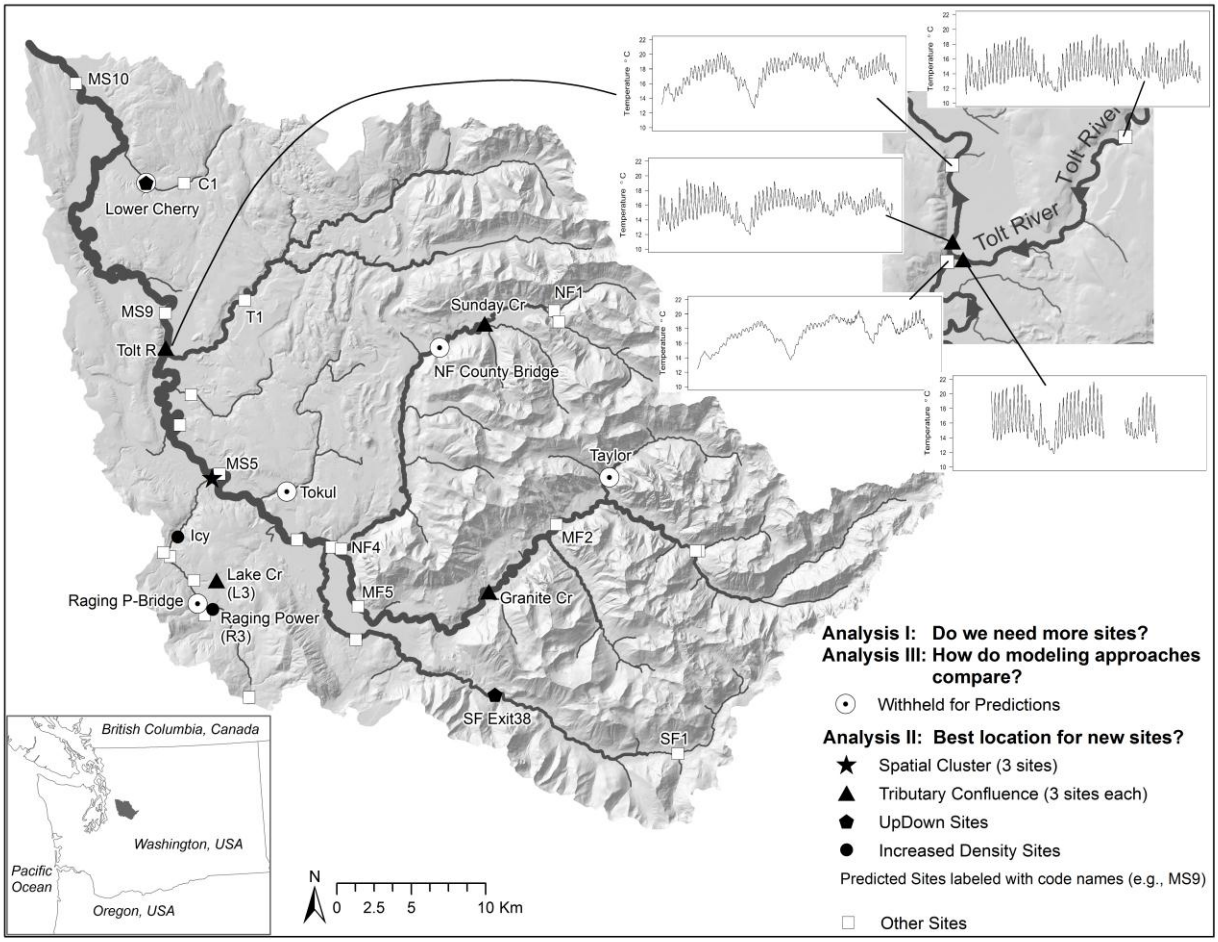
Figure 9: The model predicted metric values at the five withheld sites from SSNMs with different correlation structures and from linear models (I). SSNM correlation structures include tail-up (U), tail-down (D), Euclidean (E), combined tail-up and tail-down (UD), and combined tail-up, tail-

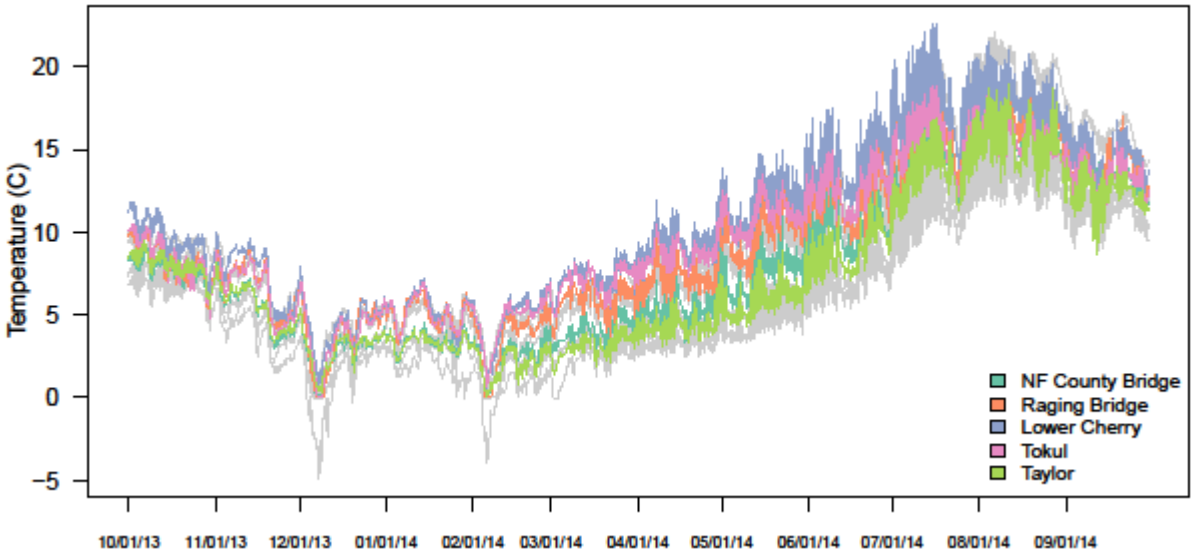
down, and Euclidean (UDE). Models were fit for four metrics (AWAT, mWAT, MWAT, NaiveVar) and during two seasons (summer and winter). All network sites except the five withheld sites were used to fit the models. Predicted values include +/- one estimated standard error and are compared to the observed metric value at each site (solid line).

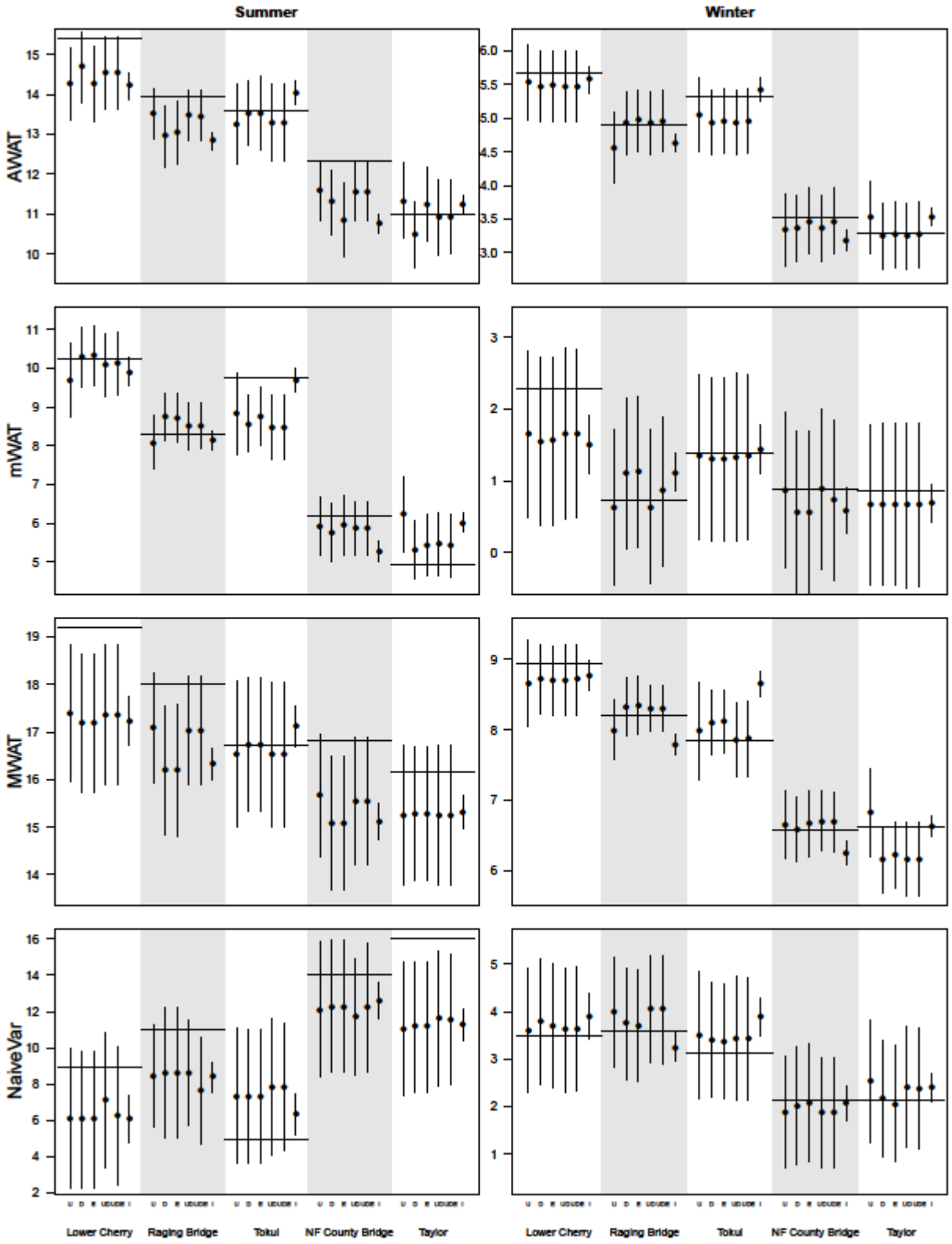
	Model Function	Season	Figure Number	# Sites Used	Metrics
I. Do we need more sites?					
	Parameter Estimation	Summer 2014	3	20-33 (ALL)	AWAT, mWAT, MWAT, NaiveVar
	Parameter Estimation	Winter 2014	3	20-34 (ALL)	AWAT, mWAT, MWAT, NaiveVar
	Prediction	Summer 2014	4	20-33	AWAT, mWAT, MWAT, NaiveVar
	Prediction	Winter 2014	4	20-34	AWAT, mWAT, MWAT, NaiveVar
II. Best location for new sites?					
	Parameter Estimation	July – Aug. 2014	Sup.	31, 33	Mean, Max, Min, NaiveVar
	Parameter Estimation	Jan. – Feb. 2015	Sup.	33, 35	Mean, Max, Min, NaiveVar
	Prediction	July – Aug. 2014	5, 6	31, 33	Mean, Max, Min, NaiveVar
	Prediction	Jan. – Feb. 2015	Sup.	33, 35	Mean, Max, Min, NaiveVar
III. How do modeling approaches compare?					
	Parameter Estimation	Summer 2014	7	20, 33	AWAT, mWAT, MWAT, NaiveVar
	Parameter Estimation	Winter 2014	7	20, 34	AWAT, mWAT, MWAT, NaiveVar
	Prediction	Summer 2014	8, 9	20, 33 (ALL)	AWAT, mWAT, MWAT, NaiveVar
	Prediction	Winter 2014	8, 9	20, 34 (ALL)	AWAT, mWAT, MWAT, NaiveVar

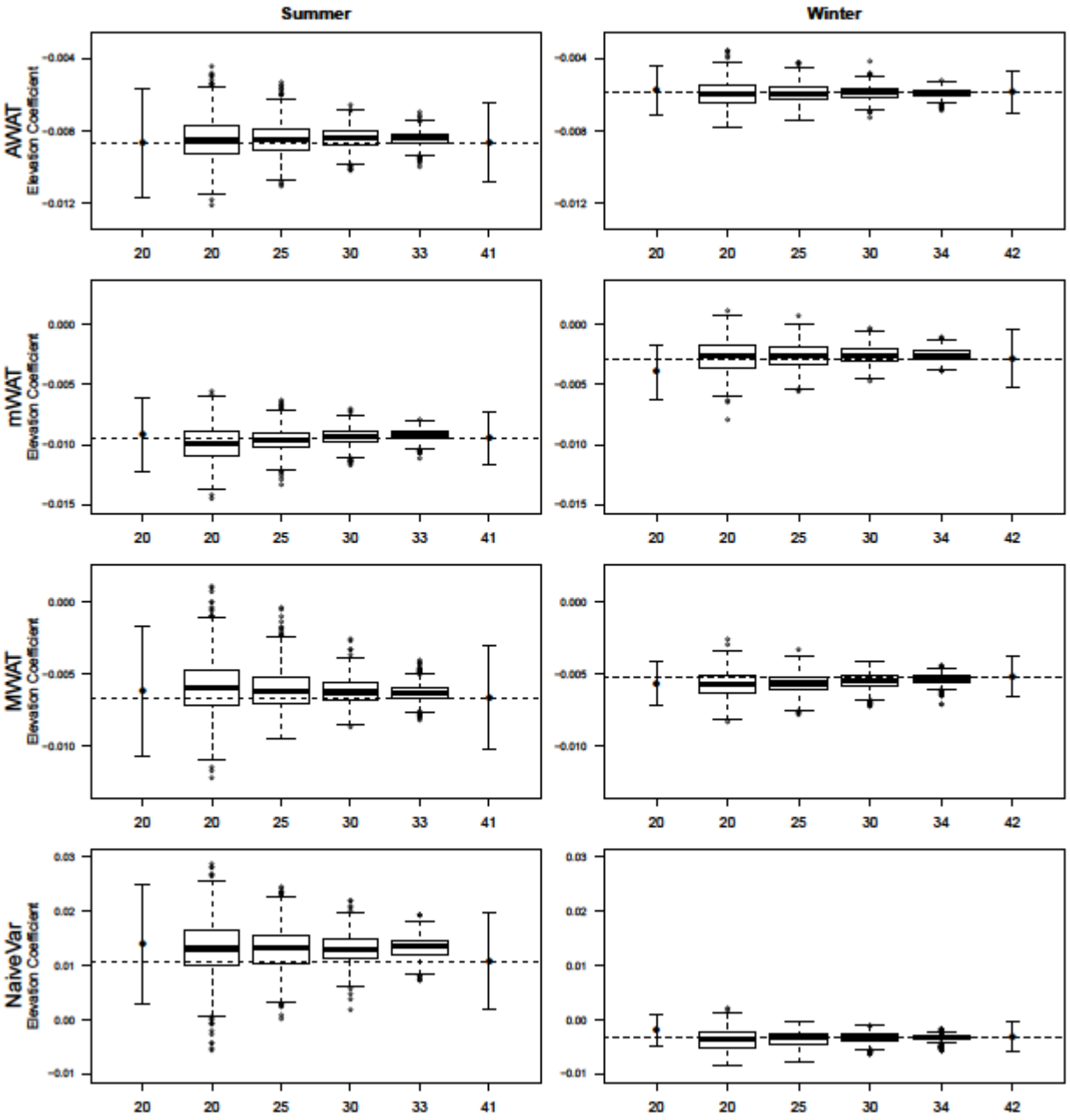
*All = 41 sites in summer and 42 sites in winter; indicates that no random sampling was used.

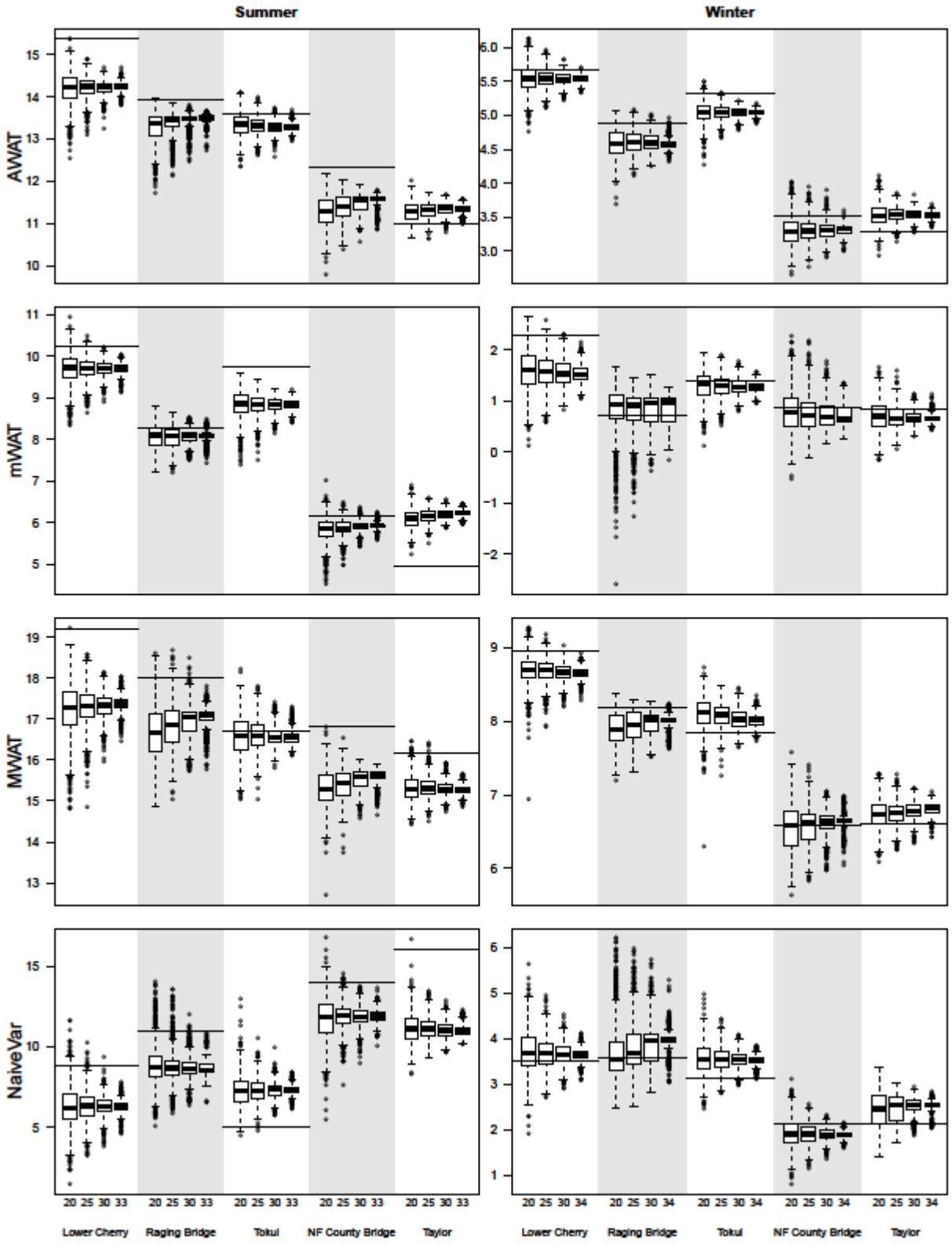
Table 1: Summary of the three analyses including model function, temporal window, associated figures, number of sites used in the random sampling or model-fitting, and metrics considered.

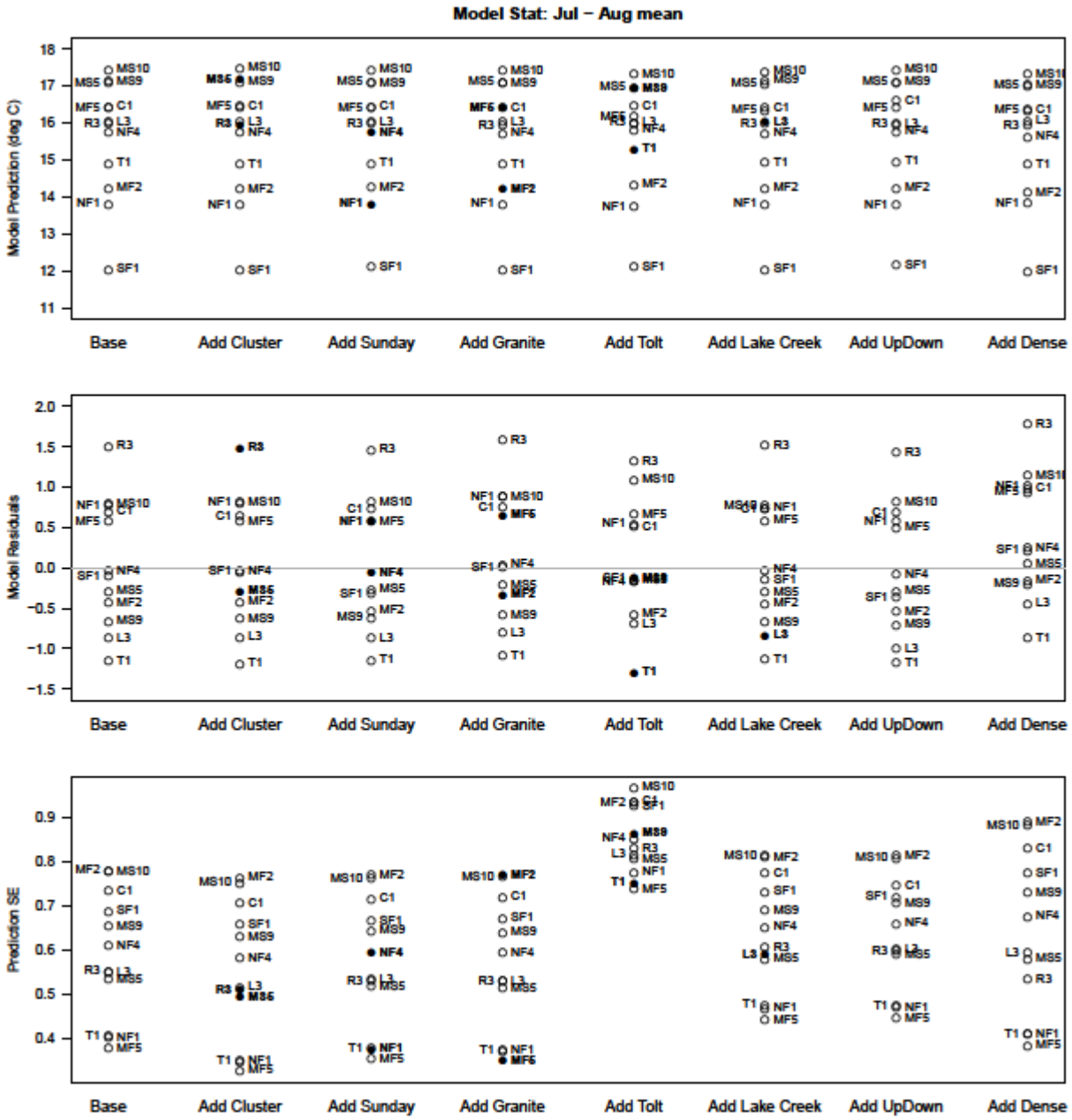

MS10
C1
Lower Cherry
MS9
T1
Tolt R
Sunday Cr
NF1
NF County Bridge
MS5
Tokul
Taylor
Icy
NF4
MF2
Lake Cr
(L3)
Raging P-Bridge
Raging Power
(R3)
MF5
Granite Cr
SF Exit38
SF1
Tolt River
Tolt River
British Columbia, Canada
Washington, USA
Pacific
Ocean
Oregon, USA
N
0 2.5 5 10 Km
Analysis I: Do we need more sites?
Analysis III: How do modeling approaches compare?
Withheld for Predictions
Analysis II: Best location for new sites?
Spatial Cluster (3 sites)
Tributary Confluence (3 sites each)
UpDown Sites
Increased Density Sites
Predicted Sites labeled with code names (e.g., MS9)
Other Sites

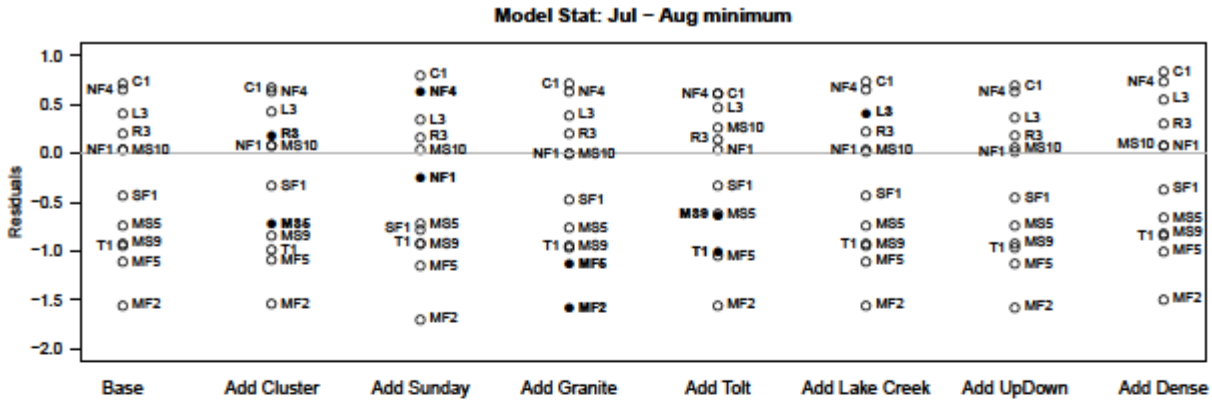

Temperature (C)
20
15
10
5
0
−5
10/01/13
11/01/13
12/01/13
01/01/14
02/01/14
03/01/14
04/01/14
05/01/14
06/01/14
07/01/14
08/01/14
09/01/14
NF County Bridge
Raging Bridge
Lower Cherry
Tokul
Taylor

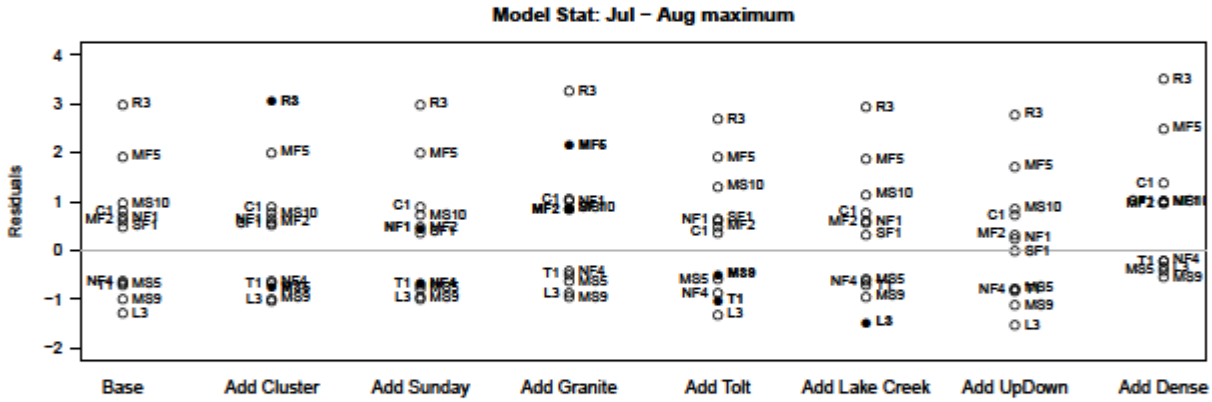

Summer
Winter
AWAT
mWAT
MWAT
NaiveVar
Lower Cherry
Raging Bridge
Tokul
NF County Bridge
Taylor

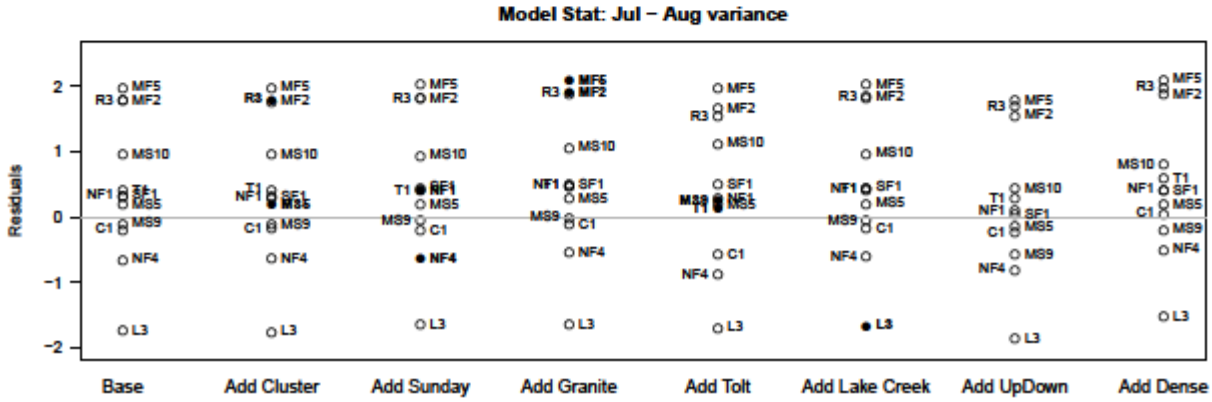

Summer
Winter
AWAT
Elevation Coefficient
mWAT
Elevation Coefficient
MWAT
Elevation Coefficient
NaiveVar
Elevation Coefficient
20
20
25
30
33
41
34
42

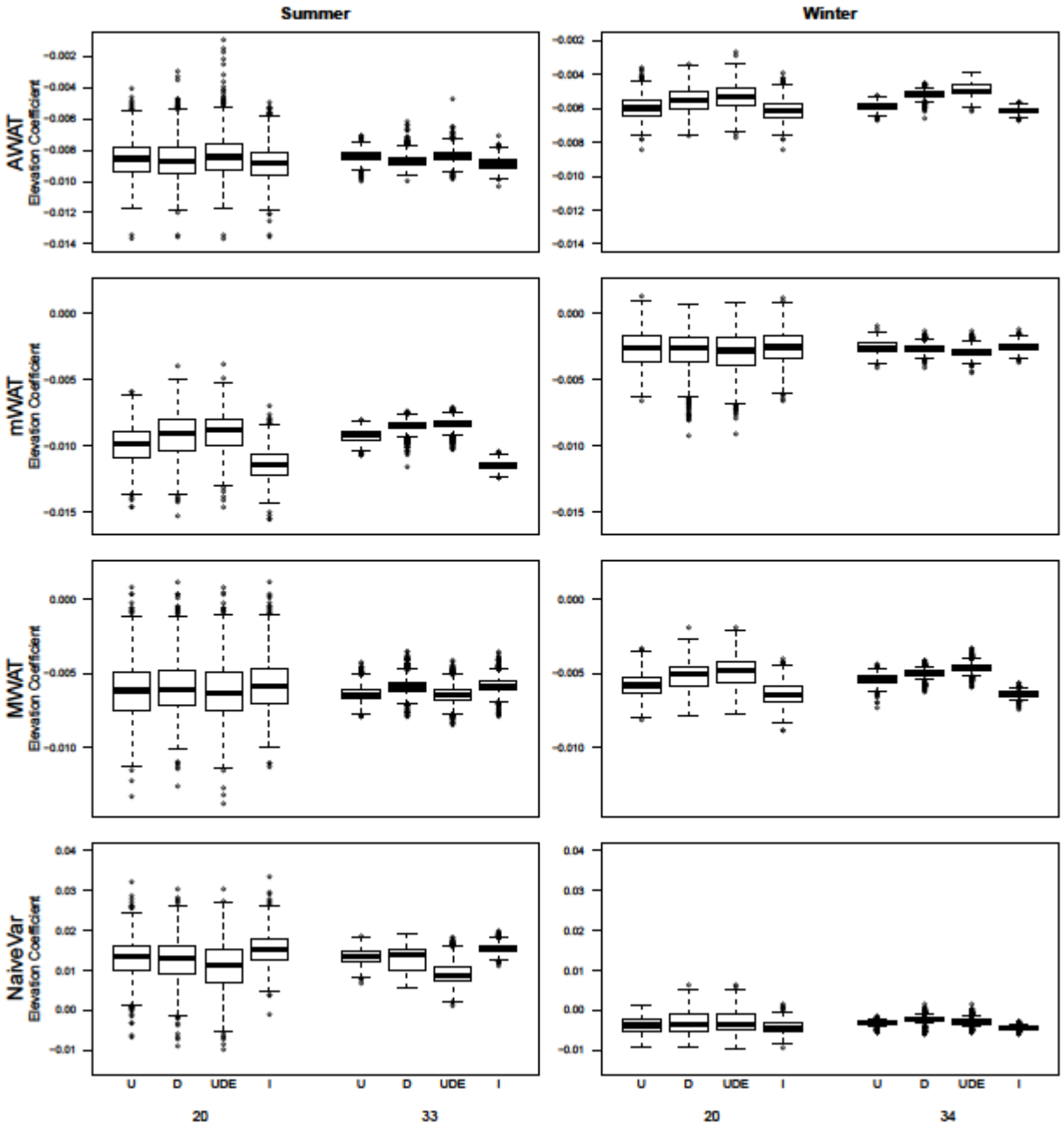

Summer
Winter
AWAT
mWAT
MWAT
NaiveVar
20 25 30 33
20 25 30 34
Lower Cherry
Raging Bridge
Tokul
NF County Bridge
Taylor

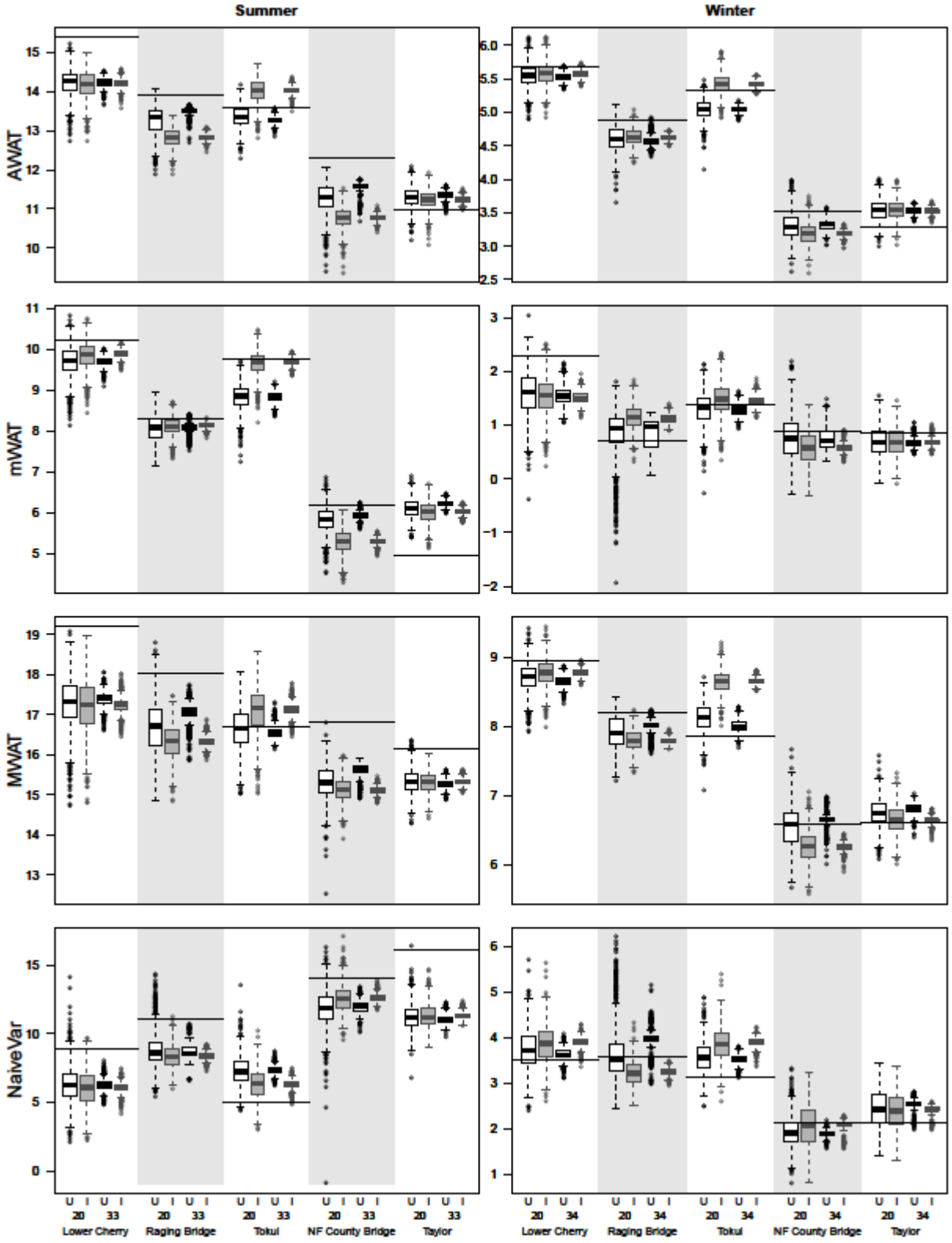

Model Stat: Jul – Aug mean
Model Prediction (deg C)
Model Residuals
Prediction SE
Base
Add Cluster
Add Sunday
Add Granite
Add Tolt
Add Lake Creek
Add UpDown
Add Dense


Model Stat: Jul – Aug minimum
Residuals
Base
Add Cluster
Add Sunday
Add Granite
Add Tolt
Add Lake Creek
Add UpDown
Add Dense


Model Stat: Jul – Aug maximum
Residuals
Base
Add Cluster
Add Sunday
Add Granite
Add Tolt
Add Lake Creek
Add UpDown
Add Dense


Model Stat: Jul – Aug variance
Residuals
Base
Add Cluster
Add Sunday
Add Granite
Add Tolt
Add Lake Creek
Add UpDown
Add Dense


Summer
Winter
AWAT
Elevation Coefficient
mWAT
Elevation Coefficient
MWAT
Elevation Coefficient
NaiveVar
Elevation Coefficient
U
D
UDE
I
20
33
20
34


Summer
Winter
AWAT
mWAT
MWAT
NaiveVar
U I U I
20 33
20 34
Lower Cherry
Raging Bridge
Tokul
NF County Bridge
Taylor