AGU PUBLICATIONS

Water Resources Research

RESEARCH ARTICLE
10.1002/2017WR020752

Key Points:
- A top-down conceptual representation of snow storage showed strong transferability across a range of challenging validation tests
- Low to moderate complexity models transfer most successfully to new conditions
- Validation-based model selection favors overfit models when calibration and validation samples are pseudoreplicated

Supporting Information:
- Supporting Information S1

Correspondence to:
C. Luce,
cluce@fs.fed.us

Citation:
Lute, A. C., & Luce, C. H. (2017). Are model transferability and complexity antithetical? Insights from validation of a variable-complexity empirical snow model in space and time. *Water Resources Research*, *53*, 8825–8850. https://doi.org/10.1002/2017WR020752

Received 13 MAR 2017
Accepted 23 SEP 2017
Accepted article online 28 SEP 2017
Published online 10 NOV 2017

Are Model Transferability And Complexity Antithetical? Insights From Validation of a Variable-Complexity Empirical Snow Model in Space and Time

A. C. Lute[1,2] and Charles H. Luce[1]

[1]United States Forest Service, Rocky Mountain Research Station, Boise, ID, USA, [2]Now at Water Resources Program, University of Idaho, Moscow, ID, USA

Abstract The related challenges of predictions in ungauged basins and predictions in ungauged climates point to the need to develop environmental models that are transferable across both space and time. Hydrologic modeling has historically focused on modelling one or only a few basins using highly parameterized conceptual or physically based models. However, model parameters and structures have been shown to change significantly when calibrated to new basins or time periods, suggesting that model complexity and model transferability may be antithetical. Empirical space-for-time models provide a framework within which to assess model transferability and any tradeoff with model complexity. Using 497 SNOTEL sites in the western U.S., we develop space-for-time models of April 1 SWE and Snow Residence Time based on mean winter temperature and cumulative winter precipitation. The transferability of the models to new conditions (in both space and time) is assessed using non-random cross-validation tests with consideration of the influence of model complexity on transferability. As others have noted, the algorithmic empirical models transfer best when minimal extrapolation in input variables is required. Temporal split-sample validations use pseudoreplicated samples, resulting in the selection of overly complex models, which has implications for the design of hydrologic model validation tests. Finally, we show that low to moderate complexity models transfer most successfully to new conditions in space and time, providing empirical confirmation of the parsimony principal.

Plain Language Summary A challenge for environmental modeling and prediction is to create models that work everywhere. This includes places where we don't have data and it also includes the future. An additional challenge is choosing an appropriate model for these new conditions.To select a model, we built snow models of varying complexity and evaluated how well they predicted snowpack in new locations and time periods. The models performed well when applied to new time periods since certain environmental variables remained constant over time, including shading and solar radiation, resulting in the time samples being statistically dependent. For these applications, validation tended to select overly complex models. When applied to new locations, the models provided good predictions as long as the conditions in the new location were not drastically different from the conditions the model was trained on. Finally, we found that simple to moderate-complexity models did better than complex models at predicting in new conditions.

1. Introduction

A core mission in hydrologic science is to improve the transferability of hydrologic models in space and time (Blöschl et al., 2013; Hrachowitz et al., 2013; McMillan et al., 2016; Montanari et al., 2013; Sivapalan et al., 2003a). Two competing narratives about how to best achieve generality are apparent in the literature: one calling for additional detail in process representation in models (the "bottom-up" approach), and a second calling for a more conceptual, scale-appropriate reframing of hydrologic modeling (the "top-down" approach) (e.g., Dooge, 1986; Hrachowitz & Clark, 2017; Liang et al., 1994; Sivapalan et al., 2003b; Young, 2003). Parsimony is a key element of this debate, usually applied to support top-down modeling approaches, and the primary lines of argument have been with respect to parameter identifiability and interpretability for either parameter transferability or diagnosis of models (e.g., Franchini & Pacciani, 1991; Jakeman & Hornberger, 1993; Kirchner, 2006; Kirchner et al., 1996; Wagener & Gupta, 2005; Wagener et al., 2001; Young, 2003). The

potential for more parsimonious models to transfer better (e.g., Burnham & Anderson, 2004; Schoups et al., 2008; Wenger & Olden, 2012) has been less well explored in hydrologic modeling.

The argument for greater complexity in models posits that poor model generality is due to a lack of process representation. Drawing on this philosophy, it is easy to conclude that another module or subroutine is needed to capture process X, Y, or Z when testing a model developed in one basin against data from another (see e.g., Kirchner et al., 1996). Following this practice to its logical conclusion is a scientific landscape filled with finely crafted models for each unique basin (e.g., McDonnell et al., 2007), accompanied by greater difficulty claiming success in transfer of these models across catchments (e.g., Blöschl et al., 2013). A critical issue is that some process representations within complex models are more conceptual than physical in their basis; consequently there can be inherent problems in transferability that can effectively only be rectified through calibration (e.g., Beven, 2001; Clark et al., 2016; Dooge, 1986; Klemeš, 1986a; McDonnell et al., 2007; Sivapalan, 2009). In other words, adding complexity may only serve to improve fits to local idiosyncratic behavior and not really serve as a means for generalization. Distinguishing whether model complexity serves generalizability or capacity to fit to individual basins can only be achieved through rigorous validation (e.g., Andréassian et al., 2009; Clark et al., 2011; Gupta et al., 2008, 2014; Klemeš, 1986b; Refsgaard & Knudsen, 1996; Seibert, 2003; Wagener et al., 2001).

These conversations highlight an empirical nature underlying much of hydrologic modeling (e.g., Gupta et al., 2012), even in "physically based" models. As we consider more top-down conceptualizations with larger lumped time and space scales, it is likely that our models will appear less "physically based" and more "empirical", but the distinction may be a misperception. Physical basis resides primarily in the conceptual representation (sensu Gupta & Nearing, 2014), which relates to choices of scales, input variables, and parameters. A key consideration is often whether the model conserves mass. If one can achieve a scale-appropriate representation in a top-down formulation, a strong physical basis can be made for model elements that are larger than plot-scale experiments. Although the hypothetical promise of physically based models is application to any location based on independent measurements of model parameters, that promise has not been widely fulfilled (e.g., Blöschl et al., 2013), and most hydrologic models still retain parameters that are not measureable outside of the context of the model itself, e.g., through calibration. The need to "transfer" information from a time and place where calibration data are available underscores the empirical aspects of hydrology.

We often conflate complexity and physical basis in hydrology because so many of our models use numerical integration algorithms operating on many spatial and/or vertical elements at sub-daily time scales. However, in an information-theoretic context, complexity refers to how much flexibility a model has when being calibrated, which is related to the number of adjustable parameters. Since we do not often assign unique adjustable parameters to the many elements in our physically based models, these models might be better termed "complicated" than "complex". An underappreciated irony is that soundly developed physically based models tend to be less complex than brute-force descriptive empiricism. An example of the latter is temperature index snowmelt models, which require several adjustable parameters to be set during calibration. In contrast, even complicated multilayer snowmelt models typically require fewer parameters, so are less *complex*, though they have greater requirements for weather input streams. Here, physical understanding can, if correctly developed and applied, constrain behaviors, limiting the need to adjust parameters. Incorrect reasoning, however, can force addition of further complexity to compensate for errors introduced by flawed logic (e.g., Mendoza et al., 2015).

To summarize these points, the goals of greater transferability, improved physical basis, and greater parsimony in models are convergent along with those of building and honing the underlying hydrologic theory (see e.g., Clark et al., 2016; Gupta & Nearing 2014). Advancing these goals is done by confronting the ideas, predictions, or models with data in a model selection process (e.g., Burnham & Anderson, 2004, Clark et al., 2011, 2015), where alternative models are calibrated then tested against new validation data from a different time and/or place.

Evaluating model transferability through model validation is particularly important in the context of climate change impact evaluation (Refsgaard et al., 2014). Uncertainty contributions from climate models alone are quite substantial (Blöschl & Montanari, 2010; Wenger et al., 2013), particularly due to high precipitation uncertainty, but additional uncertainty from hydrologic models also has consequences for recommended

adaptations or estimates of the value of mitigation (Li et al., 2012; Peel & Blöschl, 2011; Thirel et al., 2015). A specific concern for top-down models, which often seem to rely on descriptive empirical relationships, is that the underlying relationships may change over time (Hay & Clark, 2003; Lima & Lall, 2010; Merz et al., 2011; Milly et al., 2008; Wilby & Wigley, 1997), which raises questions about how well suited top-down models are for climate change impact assessments.

Changes in snow are expected to be one of the most direct and substantial effects of climate change in cold regions. While some validation of physically based snow models has been done, with models showing quite a range in performance (Essery et al., 2009; Etchevers et al., 2004; Rutter et al., 2009), little has been done for conceptual, top-down snow models such as those framed by Woods (2009) or Luce et al. (2014). Furthermore, Global Circulation Models are primarily used for assessing climatological changes and are rarely used to express changes at shorter time scales such as those typically used in physically based models (e.g., daily to sub-daily). There is consequently some utility in validating top-down models with seasonal temporal scaling.

Here we explore tradeoffs in model complexity and transferability in the context of the validation of a conceptual model of snowpack at the seasonal time scale (Luce et al., 2014). Using snowpack spatial analog (also called "space-for-time") models of varying complexity we address two primary questions: 1) how much error (a measure of transferability) should be expected when a conceptual model built on temperature and precipitation is validated under new conditions, and 2) how does the level of model complexity affect the ability of the model to transfer to new conditions? Examining these questions highlights some useful issues to consider in validation more broadly and lends support to the use of conceptual snowpack models in climate change impact studies.

2. Data and Methods

Using data from sites throughout the western U.S., we built spatial analog models of April 1 snow water equivalent (A1SWE) and Snow Residence Time (SRT). To address our first question, "how much error should be expected in validation?", we developed a transferability assessment involving two leave-one-out cross-validation (LOOCV) tests and eight non-random cross-validation (NRCV) tests. For each test we evaluated the calibration and validation statistics. We also tested how well the models predicted A1SWE and SRT in water year 2015, a particularly warm and dry winter. To address our second question, "how does model complexity affect transferability?", we built models of a range of complexities, reran the LOOCV and NRCV tests, and compared the performance of the model calibrations and validations to the level of model complexity.

2.1. Data

The spatial analog models considered two independent variables: cumulative winter precipitation and mean winter average temperature. Daily precipitation observations from Natural Resources Conservation Service (NRCS) Snowpack Telemetry (SNOTEL) sites in the contiguous western United States were obtained from http://www.wcc.nrcs.usda.gov/snow/. Sites with data for every year between water years 1991 and 2011 were selected (n = 497; locations and elevations are shown in supporting information Figure S1) and analysis was restricted to this time period to allow easy comparison with Luce et al. (2014) (see supporting information). Mean winter precipitation ($\bar{P}$) was calculated as the mean across years of the cumulative winter (November through March) precipitation at each site.

Temperature data were obtained from the TopoWx dataset (Oyler et al., 2014). The TopoWx product addresses inhomogeneities in observational data records, including step changes in SNOTEL temperature records (Oyler et al., 2015). The TopoWx SNOTEL temperature product ingests observational SNOTEL temperature data and applies United States Historical Climatology Network (USHCN) quality control methods, the pairwise homogenization algorithm of Menne and Williams (2009), and infilling procedures which consider topography and reanalysis data. We downloaded minimum and maximum TopoWx infilled temperature data for the 497 selected SNOTEL sites from ftp://mco.cfc.umt.edu/resources/TopoWx-source/auxiliary_data/station_data/. Daily mean temperatures were calculated as the arithmetic mean of TopoWx daily minimum and maximum temperatures. Mean winter temperature ($\bar{T}$) was calculated as the mean across years of the mean winter average temperature.

Two dependent variables were considered: April 1 SWE (A1SWE) and Snow Residence Time (SRT). A1SWE is commonly used in water management as an approximate metric of peak snowpack and was taken directly

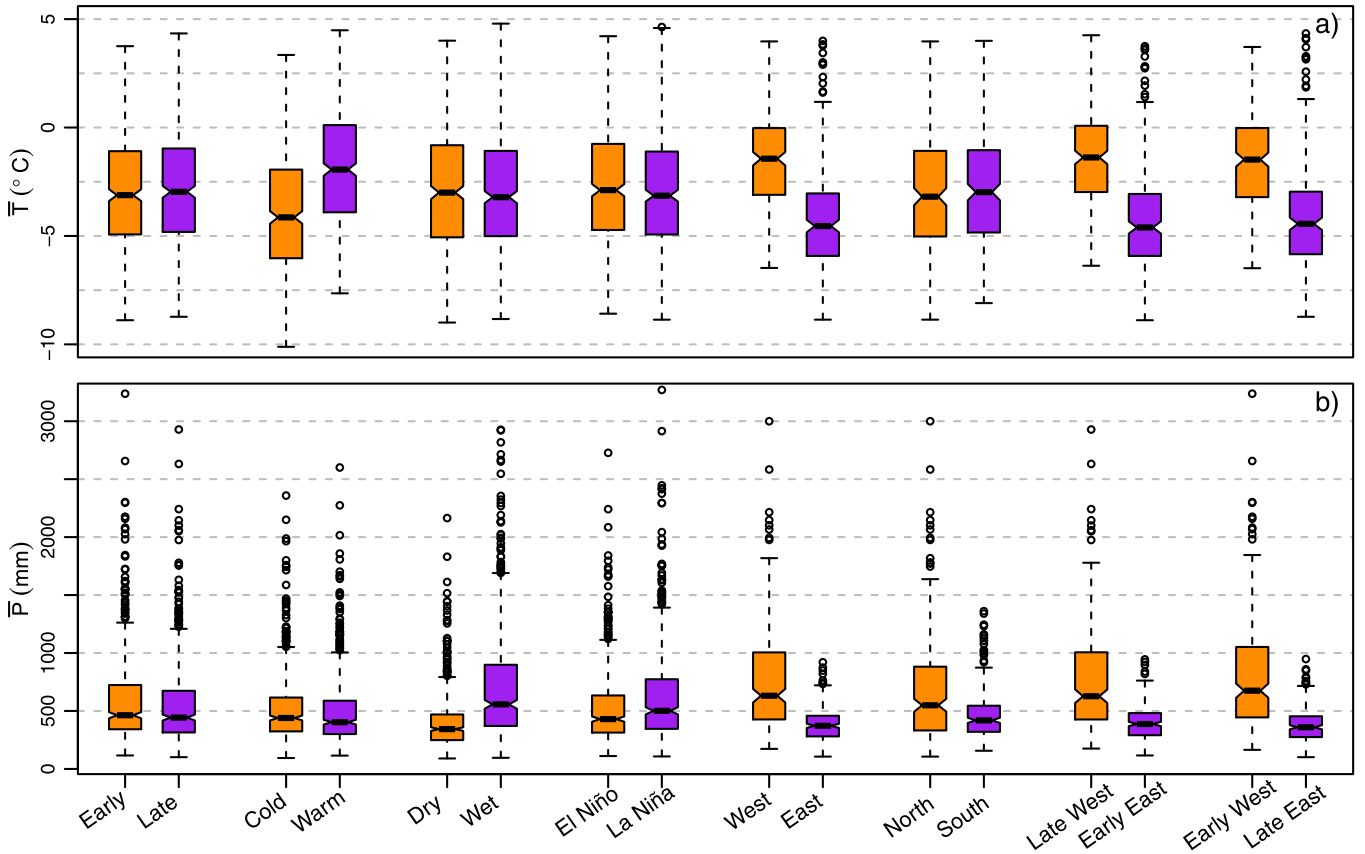

Figure 1. Distributions of a) mean winter average temperature ($\bar{T}$ (°C)) and b) mean winter cumulative precipitation ($\bar{P}$ (mm)) for each sample used in the NRCV tests. Data points are mean values across the years in the sample for a given SNOTEL site. The band in the middle of each box represents the median or 50th percentile, the bottom and top of each box represent the lower and upper quartiles. Whiskers below the box extend to the lower quartile minus 1.5 times the interquartile range (IQR) or the minimum data value, whichever is smaller. Whiskers above the box extend to the upper quartile plus 1.5 times the IQR or the maximum data value, whichever is greater.

from the SNOTEL observations. SRT is a measure of the duration of the snowpack which may be more useful to ecological and biological applications than to water management. SRT was calculated from SNOTEL records for each year as the difference between the center of timing of snow melt (*CTmelt*) and the center of timing of snow accumulation (*CTacc*), following Luce et al. (2014) (in particular note Figure 1 and equations (1) and (2) therein).

2.2. Spatial Analog Models

We used local polynomial regression of A1SWE and SRT against $\bar{T}$ and $\bar{P}$ at each site. This addresses *spatially* sampled heterogeneity in independent and response variables. Often when time series are available *temporal* sampling/modeling of such data is applied (e.g., Mote, 2006). In the context of climate change, temporal analog models consider the idea that warmer futures look like warmer years, and spatial analog models explore the idea that warmer futures resemble warmer locations. The development and interpretation of these contrasting approaches are detailed in Luce et al. (2014). A full comparison of the present models built with TopoWx temperature data and the findings of Luce et al. (2014) using original SNOTEL temperature data can be found in the supporting information.

We applied a local fit algorithm, the R locfit package (Loader, 1999, 2013), to plot a smooth surface through the data. Local regression finds a fit for the data points that resembles the unknown function $\mu(x_i)$, which describes the relationship between y (e.g., A1SWE or SRT) and $\boldsymbol{x}$ ($\bar{T}$ and $\bar{P}$):

$$y_i = \mu(x_i) + \varepsilon_i \tag{1}$$

and i indexes unique observations. Locfit does not require a priori assumptions about the form of the equation, allowing the data to direct the model including the possibility of an interaction term. In other words,

these models are algorithmic as opposed to data models (sensu Breiman, 2001). The model was specified in R as, for example:

$$\text{locfit}(\text{A1SWE} \sim \text{lp}(\overline{T}, \overline{P}, \text{nn}=0.9, \text{scale}=\text{T})) \quad (2)$$

where nn is a user defined bandwidth with values greater than zero that specifies the smoothness of the fit and how many observations (nearest neighbors) will be used for the approximation. The nn value is inversely related to the degrees of freedom of the fit. For the models presented here, nn values of 0.1, 0.5, 1.0, and 1.5 corresponded to roughly 75, 18, 7, and 6 degrees of freedom, respectively. The nn value used to fit a model can be considered a proxy for model complexity, and nn was varied according to the purposes of each experiment as described below.

2.3. Addressing Question 1: How Much Error Can Be Expected in Model Validation Under New Conditions?

Model validation entailed fitting a model to training data and then using the calibrated model to predict new data. Models were calibrated for a range of complexity levels and the complexity level with the best calibration performance was used for model validation. Drawing on the concepts outlined in Klemeš (1986b), we designed a series of validation tests of varying difficulty, ranging from leave-one-out cross-validation tests (LOOCV) to simple split-sample tests to proxy-basin differential split-sample tests. The calibration and validation performances were assessed using a range of statistical metrics detailed below.

2.3.1. Model Calibration

We framed the question of model complexity or bandwidth as a model selection problem (sensu Burnham & Anderson, 2004). We calibrated a model for all nn values from 0.1 (most complex) to 2.0 (least complex) in increments of 0.1 and noted the model with the best corrected Akaike Information Criterion (AICc) score. This range of nn values was chosen because nn = 0.1 corresponds to a high number of parameters relative to the number of sites (76 degrees of freedom for 497 sites) and the model changes very little for nn values greater than 2.

AIC is designed to assess model performance while penalizing model complexity. It is meant to be one way to estimate optimal model complexity from a calibration data set, avoiding the computational and data costs of validation (e.g., Schoups et al., 2008). Because our interest included exploring models with a large number of parameters, we used the corrected AIC, AICc, which includes corrections for circumstances when the number of data points is less than 40 times the number of parameters (Burnham & Anderson, 2004; Hurvich & Tsai, 1989). AICc is calculated as

$$AICc = n\ log\left(\hat{\sigma}^2\right) + 2K + \frac{2K(K+1)}{n-K-1} \quad (3)$$

where n is the number of observations and K is the degrees of freedom of influence of the fitted model plus two to account for the intercept and the residual variance as suggested by Burnham and Anderson (2004). $\hat{\sigma}^2$ is the maximum likelihood estimate (MLE) of residual variance:

$$\hat{\sigma}^2 = \frac{\sum \hat{\varepsilon}_i^2}{n} \quad (4)$$

where $\hat{\varepsilon}_i$ are the residuals from the fitted model. The first two terms on the right of equation (3) comprise the original AIC, with the final term on the right comprising the correction. This correction is recommended for any circumstance where one of the candidate models may have $n/K < 40$, and is used for the full suite of candidate models. AICc is a relative value with the lowest AICc among candidate models corresponding to the best model. For each test, the model complexity with the lowest calibration AICc was used in validation. The generalized cross-validation statistic (GCV) and Mallow's CP were also calculated for each calibration. GCV and CP behaved similarly to AICc.

2.3.2. Evaluation Metrics for Calibration and Validation

Model calibrations and validations were evaluated using a range of statistical metrics including the bias, root mean squared error (RMSE), and Pearson correlation r-squared value (Cor^2). Bias was calculated as the mean of the prediction errors. We also used the Nash-Sutcliffe Efficiency coefficient (NSE) which ranges from negative infinity to 1 with an ideal value at unity:

$$NSE = 1 - \frac{V_r}{V} \quad (5)$$

where V_r is the residual variance and V is the variance of the response variable. An NSE of zero indicates that the model is only as good as the mean of the values. The Kling-Gupta Efficiency coefficient (KGE) was also considered for reasons outlined in Gupta et al. (2009). The KGE is calculated as

$$KGE = 1 - \sqrt{(r-1)^2 + (\propto -1)^2 + (\beta - 1)^2} \quad (6)$$

where r is the correlation coefficient, $\propto$ is the ratio of the standard deviation of the predicted values to the standard deviation of the observed values, and β is the ratio of the mean of the predicted values to the mean of the observed values. Like NSE, KGE ranges from negative infinity to 1 and is optimized at 1.

2.3.3. Leave-One-Out Cross Validation

LOOCV tests are often used to evaluate the predictive ability of models. We built two LOOCV tests to assess the transferability of the A1SWE and SRT models. In the temporal LOOCV test, we left out one year of data and calibrated the model on the remaining years. This model was then used to predict the missing year. This was repeated for every year. The spatial LOOCV test followed a similar procedure but left out one site at a time. The model calibrated using all years at all sites (all data) is also presented for comparison.

2.3.4. Nonrandom Cross Validation

In addition to the two LOOCV tests, we developed eight NRCV tests. The tests can be thought of within the hierarchical model testing scheme developed by Klemeš (1986b), which progresses from easy tests for applications which should not challenge the model greatly (e.g., filling in a missing year of data) to more difficult tests for difficult applications (e.g., predicting in an ungauged basin). Compared to the LOOCV tests, the NRCV tests can be considered more difficult since the NRCV tests are calibrated on fewer years/sites than the LOOCV tests (i.e., fewer than 20 years/less than n−1 sites) and are required to predict more years/sites (i.e., more than 1 year or site). Each NRCV test consisted of two roughly equal sized samples of the data divided either temporally, spatially, or spatiotemporally. The tests were designed to challenge the models' ability to predict under a variety of new conditions. Each test was composed of four calculation steps comprising two calibrations and two validations: 1) the model was calibrated on sample 1 and used to predict sample 1; 2) the model was calibrated on sample 2 and used to predict sample 2; 3) the model was calibrated on sample 2 and used to predict sample 1; and 4) the model was calibrated on sample 1 and used to predict sample 2. Steps 1 and 2 are hereafter referred to as calibrations and steps 3 and 4 are referred to as validations. Throughout the paper, general tests (inclusive of all steps) are notated as, for example Cold/Warm, and particular steps are notated as, for example, Cold-Warm, indicating that the step is calibration of the model on the cold sample and prediction of the warm sample. For all samples, A1SWE, SRT, $\bar{T}$, and $\bar{P}$ were calculated as the mean value over all years in the sample at each site in the sample.

2.3.4.1. Temporal NRCV Tests

Of the eight NRCV tests, four were temporal contrasts. Following in the tradition of hydrologic model testing, the first test split the data into early and late periods, consisting of the periods 1991–2000 and 2002–2011, respectively. This test corresponds to the least challenging test in Klemeš' (1986b) hierarchy, a split-sample test, suitable for testing a model's ability to estimate data for a location it was calibrated on but at different times. The other three temporal tests can be considered differential split-sample tests, suitable for testing the models' ability to predict under changed conditions (e.g., climate change). For the second temporal test, Cold/Warm, the first sample consisted of the four years with the lowest west-wide $\bar{T}$ and the second sample was the four years with the highest $\bar{T}$. The composited sample size was decreased to four years to increase the contrast between samples and provide a more challenging test. In the third temporal test, Dry/Wet, we contrasted the four years with the highest and lowest $\bar{P}$. For the fourth temporal test, we calculated the winter average of the Multivariate El Niño-Southern Oscillation (ENSO) Index (MEI; Wolter & Timlin, 1993). Years with mean winter MEI greater than 0.4 were classified as El Niño years (n = 8), while years with mean winter MEI less than −0.4 were classified as La Niña years (n = 9).

2.3.4.2. Spatial NRCV Tests

Two of the eight NRCV tests were spatial tests (aka proxy-basin tests in Klemeš's (1986b) hierarchy) which evaluate the models' ability to predict in new locations. The first spatial test divided the sites into West (n = 247) and East (n = 250) samples along the −112.5° meridian. The second spatial test divided the sites into North (n = 246) and South (n = 251) samples along the 43rd parallel. Division lines were chosen to

create roughly equal sample sizes with a range of climates within each sample (supporting information Figure S1).

2.3.4.3. Spatiotemporal NRCV Tests

The final two NRCV tests combined the Early/Late temporal and West/East spatial tests. These are the most difficult tests in *Klemeš's* hierarchy, differential proxy-basin split-sample tests. For the Late West/Early East test, the first sample consisted of the 'late' years at the 'West' sites and the second sample consisted of the 'early' years at the 'East' sites. The Early West/Late East test was similar, with the first sample composed of the 'early' years at the 'West' sites and the second sample containing 'late' years at the 'East' sites.

2.3.5. Water Year 2015 Validation

As a final model transferability test, we extracted water year 2015 (1 October 2014 to 30 September 2015) A1SWE, SRT, and $\bar{P}$ data from SNOTEL records and water year 2015 $\bar{T}$ from the TopoWx dataset. We used the model calibrated on the 1991–2011 dataset to predict 2015 A1SWE and SRT. Winter 2015 was exceptionally dry and warm across most of the western U.S. and provides a good test of the models' ability to predict under conditions similar to those projected for the future (Cooper et al., 2016; Mote et al., 2016). Of the models of varying complexity calibrated on the full dataset, the best AICc was associated with the model with nn = 0.9 for A1SWE and nn = 0.3 for SRT. These model calibrations were therefore used to predict 2015 A1SWE and SRT, respectively. The space-for-time formulation is really intended to examine climatological variations (e.g., relatively long-term averages) rather than individual years, so this test potentially introduces more variability related to timing of events in the season, but the opportunity to contrast to the extreme year is a fortuitous example.

2.4. Addressing Question 2: How Does Model Complexity Affect Transferability?

To better understand how model complexity impacts model transferability, we ran each of the NRCV and LOOCV tests and varied the nn values from 0.1 (most complex) to 2.0 (least complex) in increments of 0.1 in both calibration (as before) and validation. We calculated the statistical metrics detailed above for each nn value, focusing on NSE and KGE. We noted the nn value associated with the greatest NSE and KGE scores and contrasted that with the nn associated with the best AICc from the training data. We further contrasted how NSE and KGE varied between the most complex model (nn = 0.1) and the model associated with the best training AICc.

3. Results

Paralleling the organization in the methods section, results are presented with a mind toward answering the two questions posed in the introduction. The following section 3.1 details the model calibration and validation performance for each of the LOOCV and NRCV tests. In the second half of the results (3.2), the interaction of model complexity and model performance is examined with reference to NSE, KGE, and AICc statistics.

3.1. Question 1: Transferability Assessment

Beginning with a discussion of the independent variables, temperature and precipitation, used in each NRCV test, we show the calibration and validation performance of the A1SWE models. This is followed by a similar analysis for the SRT models. Finally, we examine the ability of the models to predict A1SWE and SRT in water year 2015.

3.1.1. Temperature and Precipitation Contrasts for Split-Samples

Mean winter average temperature ($\bar{T}$) at SNOTEL sites ranges from −10°C to +5°C and for half of the NRCV tests the temperature distributions of the sample pairs are similar (Figure 1a). The West/East, Late West/Early East, and Early West/Late East tests have the greatest temperature differences across samples with Western sites being roughly 3°C warmer on average than Eastern sites.

Across all samples, mean winter cumulative precipitation ($\bar{P}$) ranges from 100mm to 3000 mm, with a mean of 575mm (Figure 1b). The spatial and spatiotemporal NRCV tests show the greatest precipitation contrasts between sample pairs, followed closely by the Dry/Wet test. The Dry, East, South, Early East, and Late East samples have 3rd quartile values less than the median value of their sample pair and the East, Early East, and Late East samples do not have any values greater than the third quartile of their sample pair (∼1000 mm for these cases).

3.1.2. April 1 SWE

Calibration of the spatial analog model of A1SWE (using all data) based on $\bar{T}$ and $\bar{P}$ provides a strong fit to the data with correlation, NSE, and KGE all greater than 0.85 (Table 1, row 1). The performance metrics on the validations are, of course, poorer, but still strong. The spatial LOOCV is particularly strong, indicating suitability for transfer to unmeasured sites.

The eight temporal A1SWE NRCV calibrations are strong, with all correlation, NSE, and KGE values greater than 0.80 (Table 1, gray rows in Temporal section). The calibrations display model structures (Figure 2, column 1) akin to the backward 'L' structure seen in the spatial analog models of the full dataset developed by Luce et al. (2014) and the present study (see supporting information). This structure indicates how snow responds to both temperature and precipitation in combination. The figure illustrates that both cool temperatures and substantial precipitation are required for a large snowpack, whereas low snowpack can be created by either warm temperatures or low precipitation.

The eight temporal A1SWE NRCV validations are also relatively strong, with NSE values based on the lowest AICc all exceeding 0.70 and KGE values exceeding 0.73. Biases are relatively small, with the worst (Cold-

Table 1
Summary Statistics for LOOCV and NRCV Tests of A1SWE Spatial Analog Models

		Rep	Bias (mm)	RMSE (mm)	Cor^2	NSE	KGE	Calibration nn_{AIC}
LOOCV	All Data		1.9	105.0	0.86	0.86	0.89	0.9
	Leave out one year		2.4	144.0	0.81	0.78	0.80	0.9
	Leave out one site		1.5	110.0	0.85	0.85	0.88	0.9
Temporal	Early-Early		1.4	112.5	0.86	0.86	0.88	0.9
	Late-Early		18.0	121.2	0.84	0.84	0.88	0.9
	Late-Late		2.2	106.2	0.86	0.86	0.88	0.9
	Early-Late		−14.3	113.9	0.84	0.84	0.86	0.9
	Cold-Cold		2.0	93.2	0.87	0.87	0.90	0.9
	Warm-Cold		58.2	116.9	0.85	0.80	0.85	0.9
	Warm-Warm		3.4	105.9	0.81	0.81	0.84	0.9
	Cold-Warm		−70.4	129.4	0.80	0.71	0.74	0.9
	Wet-Wet		2.9	139.7	0.87	0.87	0.89	0.9
	Dry-Wet		63.5	171.9	0.86	0.80	0.83	0.7
	Dry-Dry		0.0	79.5	0.85	0.85	0.88	0.7
	Wet-Dry		−12.1	88.0	0.83	0.81	0.77	0.9
	El Niño-El Niño		1.3	107.7	0.82	0.82	0.85	0.9
	La Niña-El Niño		−2.1	115.8	0.80	0.79	0.77	0.9
	La Niña-La Niña		2.3	112.4	0.88	0.88	0.90	0.9
	El Niño-La Niña		17.3	121.3	0.87	0.86	0.93	0.9
Spatial	West-West		4.2	136.9	0.84	0.84	0.88	1.1
	East-West	*	296.2	637.7	0.43	−2.51	−0.32	0.9
	East-East		1.4	51.8	0.93	0.93	0.94	0.9
	West-East		−48.5	80.9	0.89	0.83	0.86	1.1
	North-North		0.3	99.2	0.90	0.90	0.91	0.9
	South-North	*	178.8	378.0	0.71	−0.41	0.17	0.8
	South-South		2.8	75.1	0.90	0.90	0.91	0.8
	North-South		−62.5	126.5	0.79	0.71	0.73	0.9
Spatio-temporal	Late West-Late West		2.6	138.9	0.83	0.83	0.87	1.1
	Early East-Late West	*	250.3	546.8	0.51	−1.64	−0.13	0.9
	Early East-Early East		1.6	57.1	0.92	0.92	0.94	0.9
	Late West-Early East		−46.4	82.8	0.88	0.82	0.86	1.1
	Early West-Early West		2.1	145.0	0.84	0.84	0.88	0.9
	Late East-Early West	*	356.9	761.4	0.35	−3.40	−0.48	0.9
	Late East-Late East		0.7	52.0	0.93	0.93	0.95	0.9
	Early West-Late East		−26.8	65.1	0.91	0.89	0.91	0.9

Notes: Dark gray boxes indicate calibrations, light gray boxes indicate validations. Rows in the Rep. (data representativeness) column with an * indicate that the calibration data poorly represented the range of validation values as described in section 3.1.1. Calibration statistics are presented for the model with the best calibration AICc across nn values (shown in the Calibration nn_{AIC} column). Validation statistics are presented for the same nn value as used in the corresponding calibration.

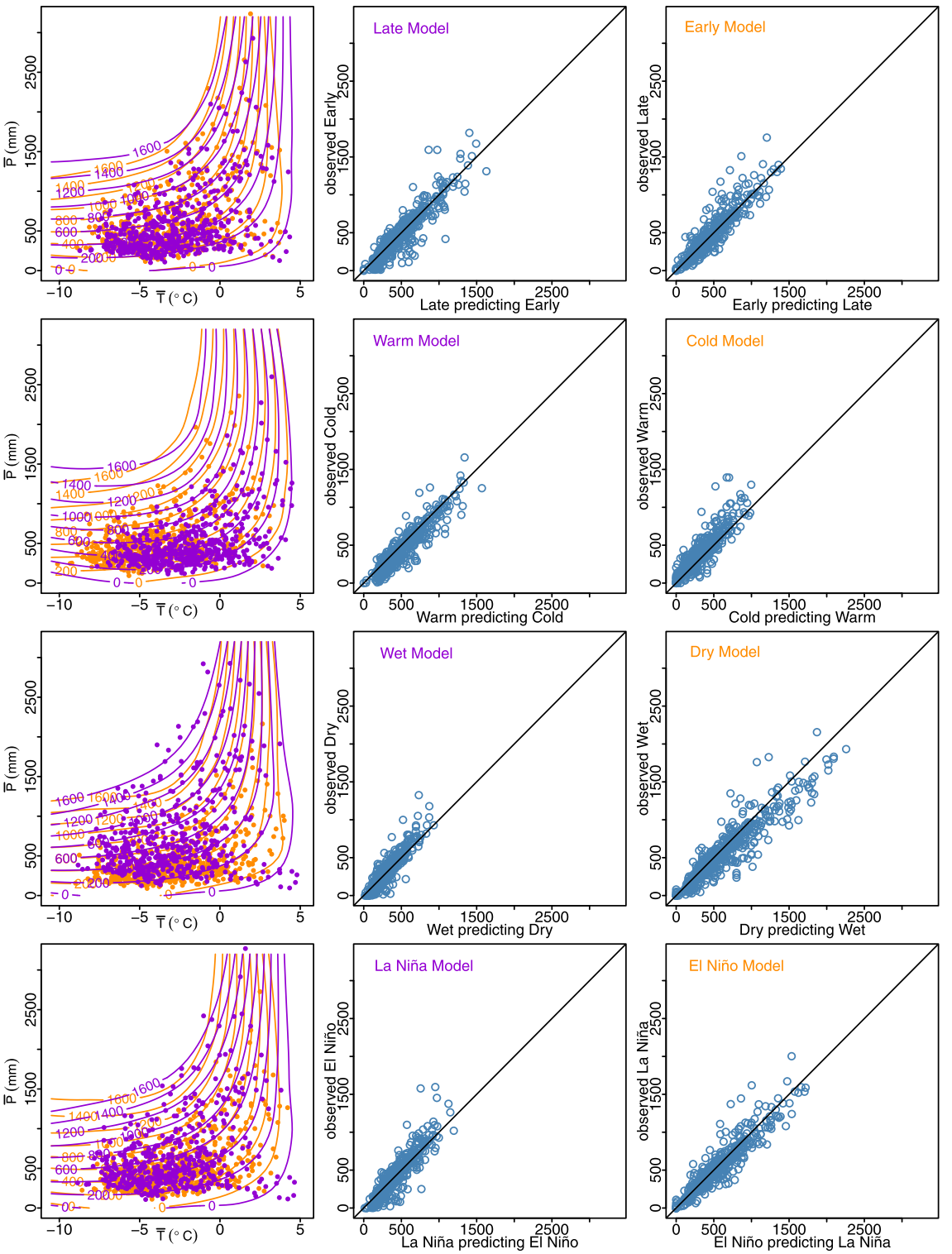

Figure 2. Summary of the temporal NRCV tests of A1SWE spatial analog models. Column 1 shows the contours of the spatial analog model calibrated on the sample dataset with the same text color in that row (e.g., orange contours in column 1, row 1 represent the spatial analog model calibrated on the Early dataset). Calibrations use the nn value of the best calibration AICc (Table 1). Contours represent gradations of A1SWE values in mm, indicated by labels on contours. The x-axis is $\bar{T}$ (°C) and the y-axis is $\bar{P}$ (mm). Points represent SNOTEL sites in $\bar{T}$-$\bar{P}$ space that are within the corresponding calibration dataset (e.g., orange points in column 1, row 1 correspond to the mean $\bar{T}$ and $\bar{P}$ values during Early years at each SNOTEL site). Columns 2 and 3 show the predicted versus observed A1SWE values (mm) for each site. Column 2 shows predictions from the purple model for the orange validation sample and column 3 shows predictions from the orange model for the purple validation sample. Black lines in columns 2 and 3 mark the 1:1 line, indicating a perfect match between predicted and observed values.

Warm, −70mm) corresponding to less than 5% of the prediction range. RMSE values are generally in the range of 110–120, less than 10% of the data range. The Dry model predicting wet conditions has with the largest RMSE in the group (172 mm), which represents about 7% of the data range.

Spatial and spatiotemporal A1SWE NRCV calibrations are strong with all correlation, NSE and KGE values greater than 0.82 (Table 1, gray rows in spatial and spatiotemporal sections). However, the calibrated models illustrate contrasting model structures, with one sample in each pair mimicking the structure seen in the temporal NRCV models and the all data model and the other sample depicting a less curved structure indicative of lower temperature sensitivity and higher precipitation sensitivity at warmer locations (Figure 3, column 1). The four calibrations resulting in less curved model structures (East, South, Early East, and Late East) lack representation of high precipitation conditions (Figure 1), paralleling issues of nonrepresentation illustrated by Seibert (2003). These models provide poor validations (Figure 3, column 2; Table 1), and greatly overestimate A1SWE in warm, wet locations such as the Cascades (not shown).

The four spatial and spatiotemporal validations for which the calibration precipitation data represented most of the range of the validation precipitation data (West-East, North-South, Late West-Early East, and Early West-Late East) are able to predict the samples with small precipitation ranges well (Table 1), although all four tend to slightly underestimate A1SWE.

3.1.3. Snow Residence Time

Calibration of the spatial analog model of SRT using all available data provides a strong fit (Table 2, row 1), although it is slightly weaker than the A1SWE model. Leave-one-out spatial validation is also strong (Table 2, row 3), indicating transferability across sites. The temporal leave-one-out validation is weaker but still satisfactory.

Calibrations of the four temporal SRT NRCV tests are strong and capture the nonlinear precipitation and temperature sensitivities present in the model with all data (Figure 4 and Table 2, gray rows in temporal section). Validations for the Early/Late NRCV tests are stronger than those for the Cold/Warm and Dry/Wet validation tests and comparable to the El Niño/La Niña validations. This may be due in part to the fact that Early/Late sample means were calculated over ten years, and the ENSO samples used eight and nine years; whereas Cold/Warm and Dry/Wet sample means were calculated over four years, making them potentially more sensitive to interannual variability in $\bar{T}$-$\bar{P}$-SRT relationships. There is also the possibility that the strong contrast in driving conditions set up by the Dry/Wet and Cold/Warm tests are more challenging, per Klemeš (1986b) expectation for differential split-sample tests.

Spatial and spatiotemporal SRT NRCV calibrations are strong with all correlation, NSE, and KGE values greater than 0.78 (Table 2, gray rows in spatial and spatiotemporal sections). Similar to A1SWE, three of the calibrations with poor precipitation representation (East, Early East, and Late East) fail to create a physically reasonable model structure for warm and wet locations (Figure 5, column 1), resulting in large underestimations of SRT in the Northern Rockies and North Cascades (not shown). Three of the four models with good data representation (West, North, and Late West) provide strong validations (Table 2). The validation using Early West to predict Late East, however, was fairly poor, and closer examination of the models (Figure 5, row 4) shows that the Early West calibration data lacks representation in a 2-dimensional sense, with a lack of samples in the warm and wet corner, yielding more significant errors for those data.

3.1.4. Water Year 2015 Validation

Water year 2015 has been highlighted as a potential example of future conditions in the western U.S. (Cooper et al., 2016; Mote et al., 2016; Sproles et al., 2017). Despite the fact that our models are built on climatological means, comparison to this unique year presents an unusual opportunity to test the model on expected 'future' conditions. Averaged across SNOTEL sites, winter 2014–2015 was 2.3°C warmer and 15% drier than the mean conditions in the calibration period (1991–2011), with regional variability including +4–5° temperature anomalies in the Pacific Northwest and 49% of normal precipitation in California. Roughly 25% of SNOTEL sites had no snow on 1 April 2015, whereas about 5% had no snow on average for the years 1991–2011 and the previous worst year had 13% of sites with no snow. Sites with no snow on 1 April 2015 were distributed across climatologically warmer areas, with only two sites in Wyoming and Colorado. Average SRT across all sites was about 27% less than the mean SRT during the calibration period.

The spatial analog model of A1SWE calibrated to the full dataset overpredicts 2015 A1SWE in some areas with observed low or zero snow (such as western Oregon and Washington), but does well for the other sites

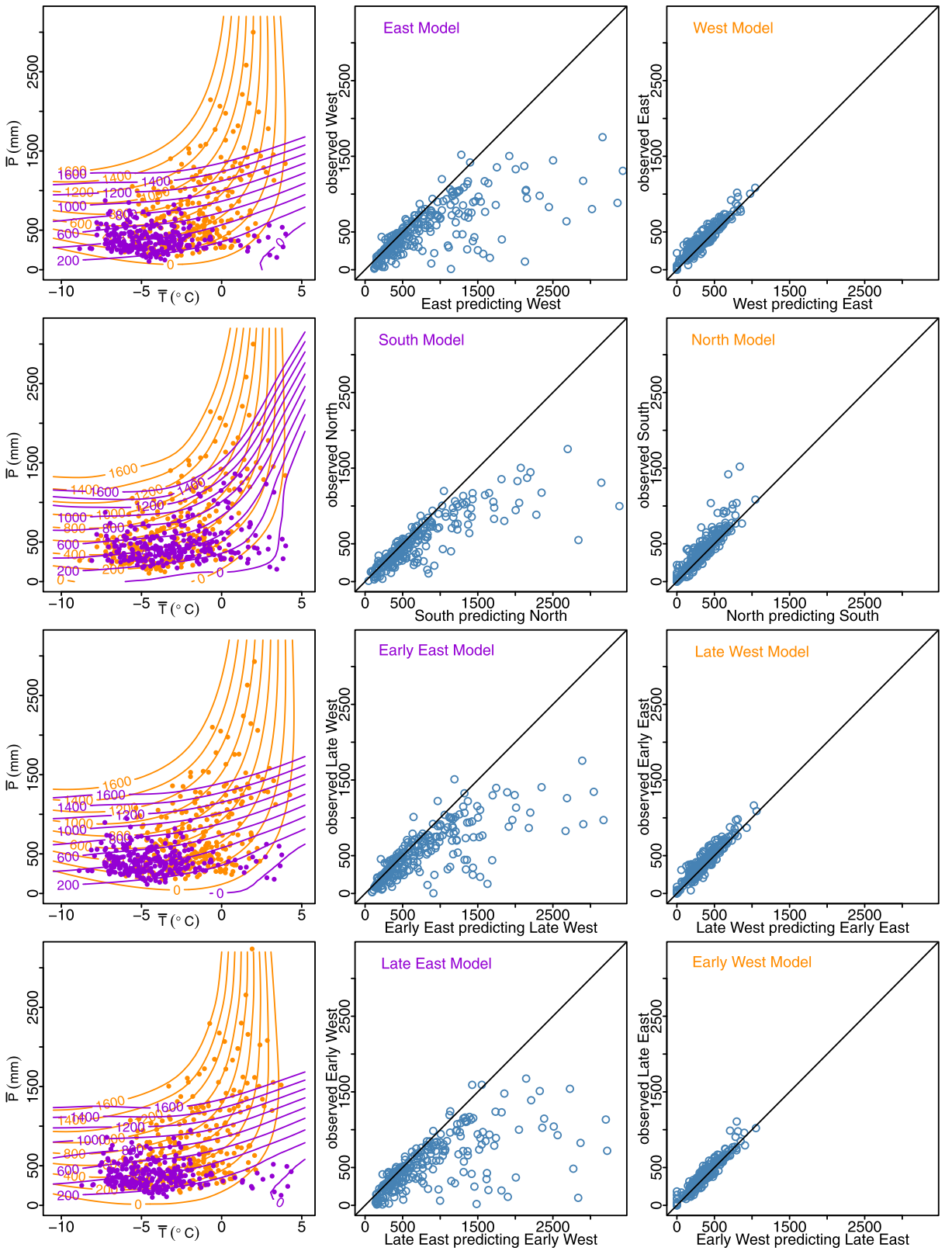

Figure 3. Summary of the spatial and spatiotemporal NRCV tests of A1SWE spatial analog models. See description of Figure 2 for explanation.

Table 2
Summary Statistics for LOOCV and NRCV Tests of SRT Spatial Analog Models

		Rep	Bias (mm)	RMSE (mm)	Cor^2	NSE	KGE	Calibration nn_{AIC}
LOOCV	All Data		0.2	9.4	0.83	0.83	0.87	0.3
	Leave out one year		−1.1	15.7	0.68	0.63	0.74	0.3
	Leave out one site		0.2	10.0	0.81	0.81	0.87	0.3
Temporal	Early-Early		0.2	9.7	0.84	0.84	0.88	0.3
	Late-Early		−1.6	10.7	0.82	0.81	0.79	0.5
	Late-Late		0.0	10.2	0.80	0.80	0.84	0.5
	Early-Late		1.8	10.8	0.79	0.77	0.88	0.3
	Cold-Cold		0.1	9.3	0.79	0.79	0.84	0.6
	Warm-Cold		6.1	12.3	0.73	0.64	0.74	0.4
	Warm-Warm		0.1	12.0	0.77	0.77	0.82	0.4
	Cold-Warm		−7.4	15.2	0.73	0.64	0.72	0.6
	Wet-Wet		0.3	11.7	0.83	0.83	0.87	0.6
	Dry-Wet		4.1	15.6	0.72	0.70	0.78	0.5
	Dry-Dry		0.0	10.9	0.76	0.76	0.80	0.5
	Wet-Dry		−4.0	13.4	0.68	0.64	0.81	0.6
	El Niño-El Niño		0.1	10.3	0.80	0.80	0.83	0.5
	La Niña-El Niño		−3.1	11.4	0.78	0.76	0.79	0.5
	La Niña-La Niña		0.2	9.8	0.84	0.84	0.87	0.5
	El Niño-La Niña		3.2	11.0	0.82	0.80	0.86	0.5
Spatial	West-West		0.2	10.1	0.84	0.84	0.88	0.8
	East-West	*	−6.2	19.5	0.48	0.41	0.62	1.0
	East-East		0.2	8.6	0.82	0.82	0.86	1.0
	West-East		5.8	13.0	0.78	0.60	0.76	0.8
	North-North		0.1	8.9	0.86	0.86	0.90	0.7
	South-North	*	−0.3	13.4	0.69	0.68	0.80	0.9
	South-South		0.1	9.1	0.81	0.81	0.85	0.9
	North-South		0.4	13.3	0.77	0.60	0.69	0.7
Spatio-temporal	Late West-Late West		0.2	11.2	0.79	0.79	0.84	0.8
	Early East-Late West	*	−3.5	16.1	0.60	0.58	0.68	1.3
	Early East-Early East		0.2	9.2	0.82	0.82	0.85	1.3
	Late West-Early East		3.3	11.3	0.79	0.72	0.84	0.8
	Early West-Early West		0.3	10.1	0.86	0.86	0.90	0.9
	Late East-Early West	*	−10.0	24.6	0.36	0.18	0.54	1.0
	Late East-Late East		0.0	8.6	0.82	0.82	0.87	1.0
	Early West-Late East		8.6	16.0	0.72	0.38	0.68	0.9

Note: Same as Table 1, except for SRT instead of A1SWE.

(overall bias = 47.5mm, RMSE = 124mm, Cor^2 = 0.72, NSE = 0.66, KGE = 0.72; Figure 6a). The SRT model predicts 2015 SRT well (bias = −4.5 days, RMSE = 19.0 days, Cor^2 = 0.70, NSE = 0.66, KGE = 0.67; Figure 6b), but shows a bias in slope compared to the 1:1 line rather than a bias in the mean values, which is not directly flagged in our metrics. Specifically, there was a tendency to underpredict SRT at sites with long SRT (such as is common at interior west sites) and overpredict SRT in locations with short SRT (such as is commonly seen in western Washington and Oregon).

3.2. Question 2: Transferability Versus Complexity

The question of how the level of complexity affects transferability has two components. We begin by examining which models are more transferable, complex versus parsimonious, in the sense of a model selection exercise. We then ask if there is any difference in performance as we look at simpler models.

3.2.1. What Level of Model Complexity Is Most Transferable?

In all model calibrations, the most complex model formulation (nn = 0.1) has the highest NSE and KGE values (Tables 3 and 4). More parameters generally allow for a closer fit in calibration. In contrast, the best AICc values are associated with moderate complexity models in calibration, ranging from nn = 0.7 to 1.1 for A1SWE and nn = 0.3 to 1.3 for SRT. The AICc includes a penalty for more parameters, supporting the selection of less complex models.

The nn values that would produce the optimal NSE in validation of A1SWE are 1) usually larger than 0.1, typically in the 0.7 to 0.9 range (low to moderate complexity), and 2) commonly comparable to that selected

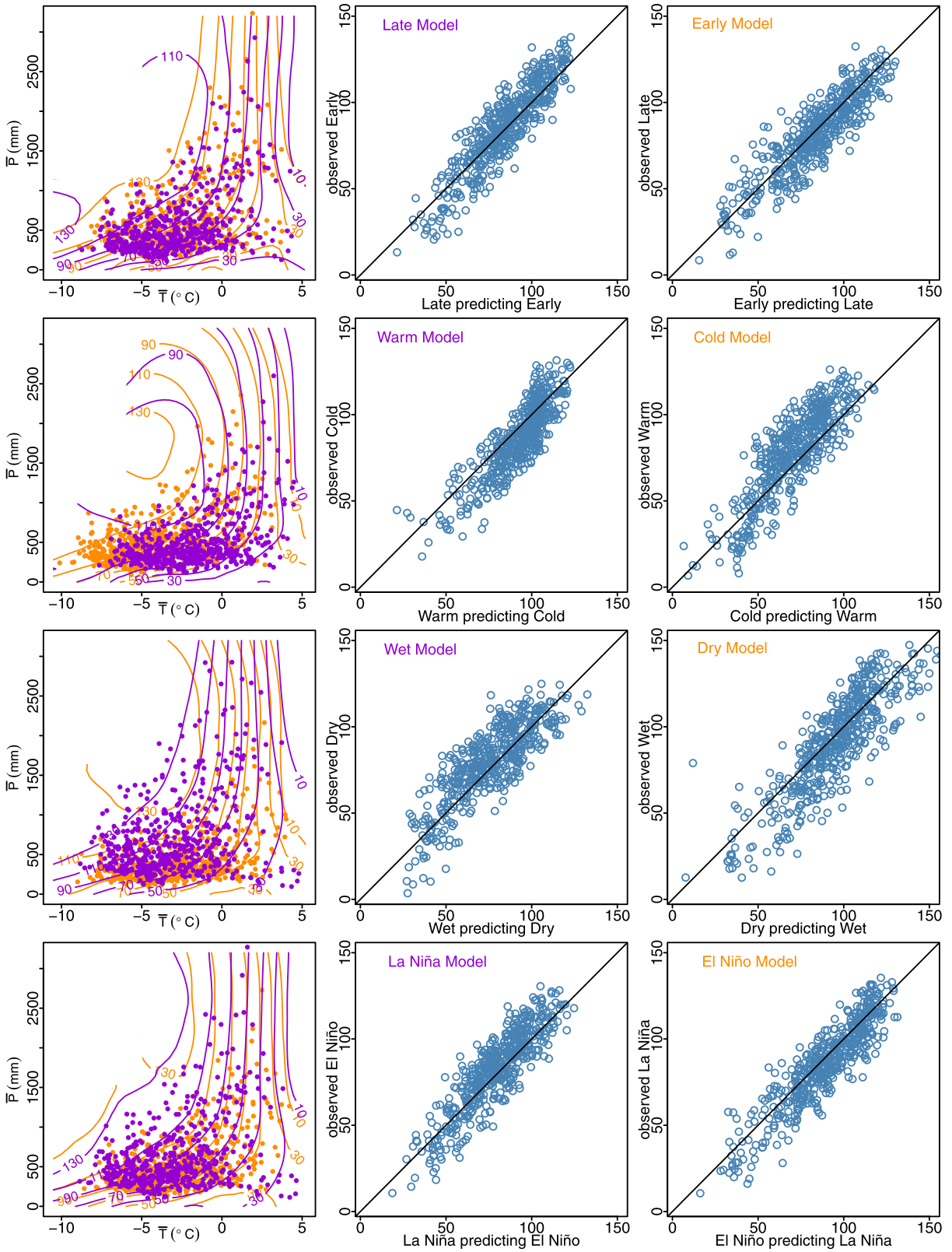

Figure 4. Summary of the temporal NRCV tests of SRT spatial analog models. See description of Figure 2 for explanation. Units of SRT are days.

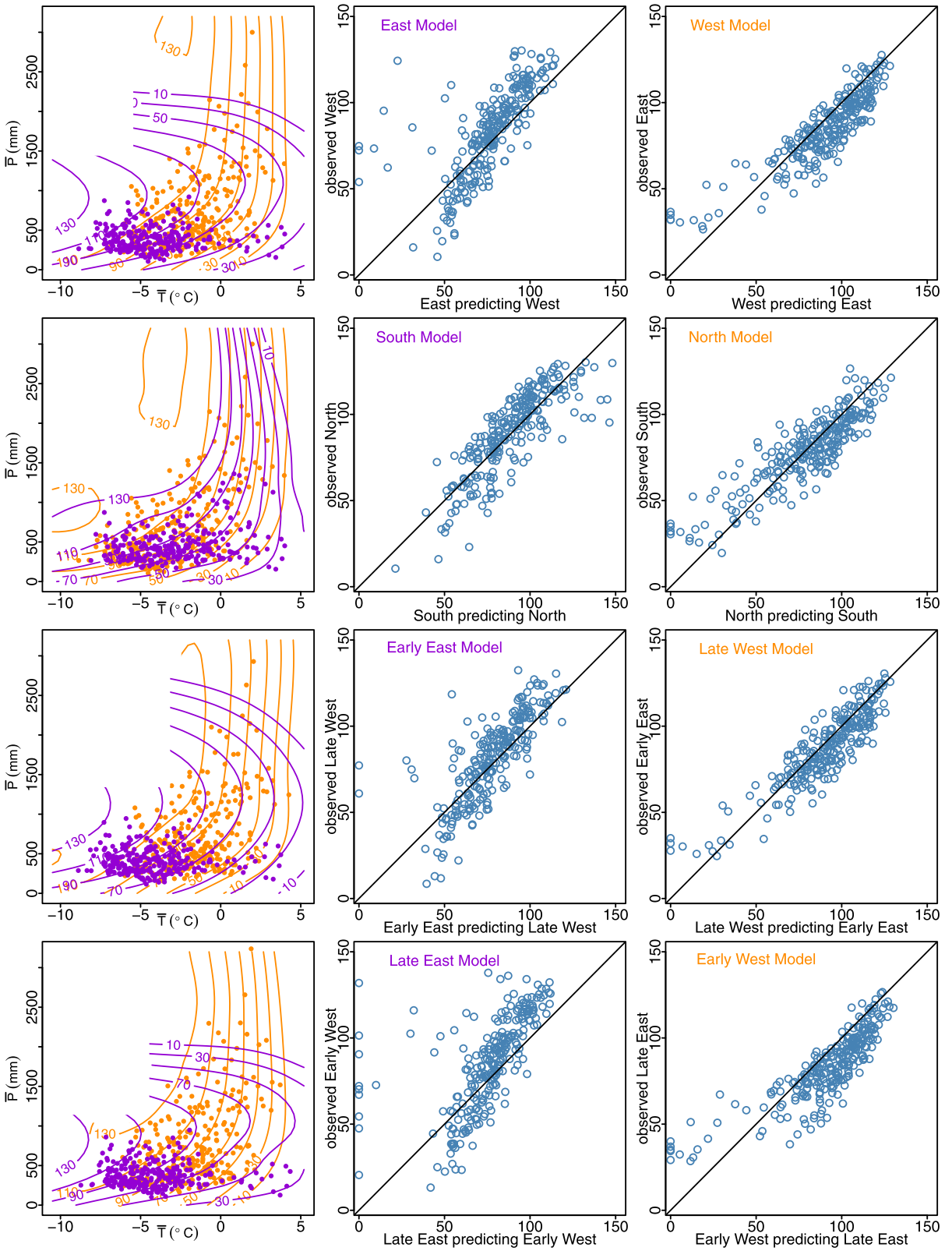

Figure 5. Summary of the spatial and spatiotemporal NRCV tests of SRT spatial analog models. See description of Figure 2 for explanation. Units of SRT are days.

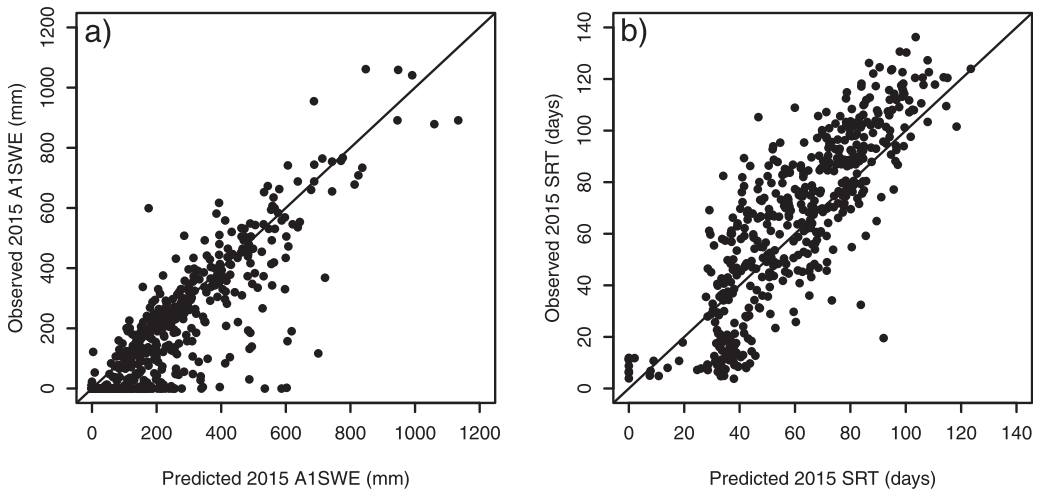

Figure 6. Predicted versus observed winter 2015 a) A1SWE (mm) and b) SRT (days). Each point represents a SNOTEL site. Black lines mark the 1:1 line or perfect correspondence between predictions and observations. Predictions are from spatial analog models calibrated on the 1991–2011 period using the nn value of the best calibration AICc from the all data model (nn = 0.9 for A1SWE and nn = 0.3 for SRT).

by AICc (Figure 7). Similar patterns are noted for KGE based model selection (Table 3). Deviations from these generalizations have likely causes in two cases. First, these generalizations do not apply to the four validations where the calibration data did not represent the range of the validation data well. In general, the models with poor data representation fail to capture key aspects of the model structure and therefore the complexity of the most transferable validation model is irrelevant. Second, the Early/Late and ENSO pairs select higher complexity (low nn) models in validation compared to their AICc suggestions and to the nn selections for other pairs.

The patterns for SRT were less consistent than for A1SWE, but the basic guidelines were similar (Table 4). nn values for optimal NSE were generally greater than 0.1 but ranged from 0.3 to 1.2. AICc selected models in the range of nn = 0.3 to 0.6 for the temporal models and 0.7 to 0.9 for spatial and spatiotemporal models with representative calibration data. Again, the Early/Late validations had the lowest NSE-selected nn value (the Early-Late validation selected the most complex model).

The ability of the most complex model formulation to perform best for the Early/Late split-sample validations implies that any idiosyncrasies present in the calibration data are also present in the validation data, so that the validation data look very much like the calibration data. This is confirmed by the statistically significant correlations between the residuals from the temporal split-sample calibrations with nn = 0.1. The residuals from the Early/Late calibrations are strongly correlated (A1SWE: R^2=0.79, $p < 1\times10^{-15}$; SRT: R^2=0.64, $p < 1\times10^{-15}$, Figure 8). Correlated calibration errors are also present for the El Niño/La Niña (R^2=0.72), Dry/Wet (R^2=0.58), and Cold/Warm (R^2=0.76) April 1 SWE split samples, but the strength of local idiosyncrasies is lower, presumably because the contrast between the two data sets overrides the amount of variability accounted for in a more complex model. The presence of correlated calibration errors means that these splits are not completely independent samples but pseudoreplications (sensu Hurlbert, 1984), and if the idiosyncratic variations outweigh the temporal contrasts, then validations will tend to favor complex models. A ratio of the difference in mean SWE to the average RMSE of the calibrations is low (0.16) for the Early/Late pair and high (2.28) for the Dry/Wet pair, reflecting the strong difference in selecting complex versus simpler models. The ratios are more intermediate for the El Niño/La Niña (0.98) and Cold/Warm (0.58) split samples, which is reflected in the El Niño/La Niña selection of a somewhat higher complexity model than suggested by AICc, but is not as strongly reflected in the Cold/Warm selection of a model of complexity comparable to the AICc selection.

3.2.2. How Much Better Do Less Complex Models Do in Validation?

Figure 9 shows NSE values for April 1 SWE and SRT validations. In general, models selected by AICc in calibration (purple) are relatively strong models of April 1 SWE (NSE = 0.7 to 0.9, Figure 9a), and these are generally comparable to (but slightly weaker than) the model associated with the best NSE found in validation (orange). The contrast between the most complex model and the model with the best AICc is generally

Table 3
A1SWE Model Complexity Versus Transferability

		Rep	nn_{NSE}	nn_{KGE}	nn_{AIC}
LOOCV	All Data		0.1	0.1	0.9
	Leave out one year		0.8	0.2	
	Leave out one site		0.9	0.3	
Temporal	Early-Early		0.1	0.1	0.9
	Late-Early		0.1	0.1	
	Late-Late		0.1	0.1	0.9
	Early-Late		0.1	0.1	
	Cold-Cold		0.1	0.1	0.9
	Warm-Cold		0.7	1.1	
	Warm-Warm		0.1	0.1	0.9
	Cold-Warm		0.9	0.6	
	Wet-Wet		0.1	0.1	0.9
	Dry-Wet		1.3	1.2	
	Dry-Dry		0.1	0.1	0.7
	Wet-Dry		0.8	0.2	
	El Niño-El Niño		0.1	0.1	0.9
	La Niña-El Niño		0.3	0.1	
	La Niña-La Niña		0.1	0.1	0.9
	El Niño-La Niña		0.4	0.4	
Spatial	West-West		0.1	0.1	1.1
	East-West	*	0.1	0.1	
	East-East		0.1	0.1	0.9
	West-East		0.9	0.8	
	North-North		0.1	0.1	0.9
	South-North	*	0.1	1.0	
	South-South		0.1	0.1	0.8
	North-South		1.1	1.2	
Spatio-temporal	Late West-Late West		0.1	0.1	1.1
	Early East-Late West	*	0.1	0.1	
	Early East-Early East		0.1	0.1	0.9
	Late West-Early East		0.8	0.8	
	Early West-Early West		0.1	0.1	0.9
	Late East-Early West	*	0.1	0.1	
	Late East-Late East		0.1	0.1	0.9
	Early West-Late East		0.9	0.8	

Notes: For each A1SWE LOOCV and NRCV test, the nn values of the models with the best NSE, KGE, and AICc, respectively are shown. * in the Rep. (data representativeness) column indicate that the calibration data poorly represent the range of values in the validation data.

greater for A1SWE models than for SRT models. None of these generalizations apply where calibration data are not representative. The AICc-selected models are generally stronger than the most complex model, except in the case of the early versus late pair, where we suspect pseudoreplication may influence the results. The improvement in model selection using AICc or validation is most pronounced for the spatial comparisons. General patterns are similar for KGE (Table 5).

For SRT models, validations are weaker but still acceptable (NSE = 0.6 to 0.8, Figure 9b) in all but one case (ignoring the validations with non-representative data). The Early West model dramatically underpredicts the persistence of Late East snow (Figure 5, row 4, column 3) leading to large squared errors. While the univariate representativeness of Early West data is similar to that of the Late West data (Figure 1), in a bivariate sense, all of the warm points in the Early West data also have high precipitation (Figure 5, row 4, column 1), and there is a lack of data for fitting warm and low precipitation conditions. This particular lack has greater consequences for the SRT model than the April 1 SWE model.

4. Discussion

Spatial analog models of April 1 SWE and Snow Residence Time transferred well across time and space, with a few considerations. In particular, the models transferred less successfully across conditions that required substantial extrapolation. Less complex models transferred better than the most complex models, in some cases with substantial improvement in NSE.

4.1. How Well Did the Models Transfer?

Although the models generally transfer well to new conditions in both space and time, the transferability assessment brought to light three considerations that may be applicable to model validation more generally: 1) extrapolation, 2) pseudoreplication, and 3) degree of physical basis. We address each of them in turn.

4.1.1. Extrapolation

Validation performance varied with the degree of contrast between calibration and validation climates (in terms of temperature and precipitation). This finding for snow models agrees with similar findings for rainfall-runoff models. Seibert (2003) demonstrated the poor ability of a model calibrated on low flows to predict high flows and noted that the decrease in performance between calibration and validation was greater when the difference between the datasets was greater. Using three rainfall-runoff models, Coron et al. (2012) also showed that validation error increased as the contrast in precipitation between calibration and validation increased. These findings are important for any application of a hydrologic model to contrasting conditions, including forecasting, simulating extreme events, modeling ungauged basins, uncertainty estimations, and of course, climate change impacts evaluations. Spatial analog models expand the limits imposed by extrapolation by incorporating a greater diversity of climates in the calibration dataset, which may improve the quality of hydrologic predictions under new conditions, including a changing climate (Singh et al., 2011). We particularly noted that strong temporal contrasts (Dry/Wet and Cold/Warm) still transferred well despite large shifts in the central tendency because the large spatial sample enabled some representation across much of the target data range within the outliers of each training set (Figure 2).

The use of a variety of NRCV tests as opposed to a single validation provides a more nuanced understanding of the extrapolation capacity of the models. Understanding the ability of models to predict under conditions that are significantly different than those they were calibrated on is of vital importance to the twin problems of Predictions in Ungauged Basins and Predictions in Ungauged Climates. Spatial and spatiotemporal NRCV tests can

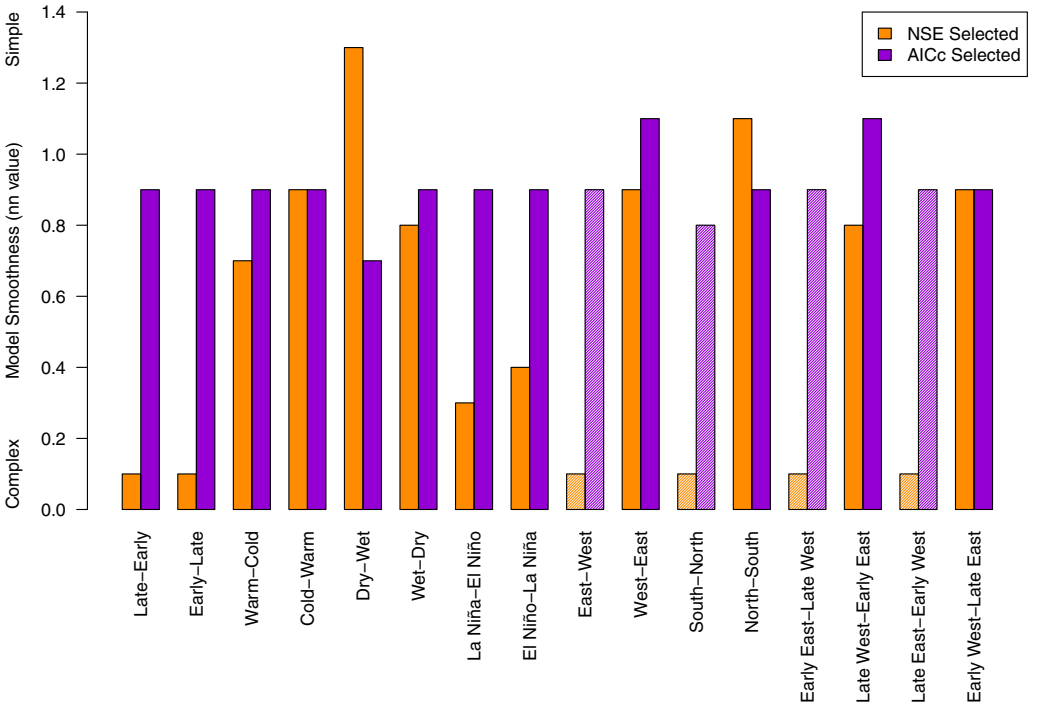

Figure 7. Comparison of nn selected by the best NSE value in validation (orange) and the nn selected during calibration by AICc (purple) for the April 1 SWE models. Validation pairs with poor representation in the calibration data are highlighted with cross-hatching.

provide some indication of how well the model will perform under new conditions (Klemeš, 1986a, 1986b) and the use of multiple validations adds nuance to his understanding (see e.g., Coron et al., 2012). Given the tremendous uncertainties associated with hydrologic prediction in a changing climate, understanding the extrapolation capacity of the models used to make the predictions is essential.

4.1.2. Pseudoreplication

Our results showing that the validation of the spatial model between the Early and Late time periods selected the most complex model may have important implications for more traditional hydrologic model validation. The problem is termed pseudoreplication by Hurlbert (1984), although the consequences we wish to discuss are different than the inference problems he considered. While our "experiment" has substantial spatial replication, there is no true replication of the experiment in time. Year after year, we use the same "plots" with particular equipment, wind exposure, shading, and other local idiosyncrasies in a single repeated measures experiment. Each site's annual variability is small relative to the larger domain of temperature and precipitation (Figure 10), and this is even truer when temporal averages are taken. Consequently, each site affects the same small region of the paired models in a comparison unless temporal splits have strong contrasts ensuring that a given station sits in a fairly different place in $\bar{T}$-$\bar{P}$ space. Analysis methods exist for inference from repeated measures (e.g., correction of p-values), but there is less awareness of the consequences of using repeated measures for model selection through model validation. In particular, we note that validation using repeated measures sampling can lead to overfitting.

In the current case we have a model that generalizes across a large number of spatial samples, and we wish to demonstrate that a calibration in one time period applies to (transfers to) a different time period. The model that generalized best through time in the Early/Late experiment, using every station, was the *most complex* spatial model. This model preserved unique behaviors caused by persistent local biases across time (Figure 8). In contrast, any validation applied to stations other than those involved in the calibration (e.g., the spatial and spatiotemporal NRCV tests) and validations with strong temporal contrasts selected models of low-intermediate complexity.

Flipping the problem on its side, consider an example where a model is built to generalize across time in one place, a common sort of model in hydrology. While something can be learned of hydrology in general

Table 4
SRT Model Complexity Versus Transferability

		Rep	nn_{NSE}	nn_{KGE}	nn_{AIC}
LOOCV	All Data		0.1	0.1	0.3
	Leave out one year		0.8	0.3	
	Leave out one site		0.3	0.3	
Temporal	Early-Early		0.1	0.1	0.3
	Late-Early		0.3	0.1	
	Late-Late		0.1	0.1	0.5
	Early-Late		0.1	0.1	
	Cold-Cold		0.1	0.1	0.6
	Warm-Cold		0.5	1.5	
	Warm-Warm		0.1	0.1	0.4
	Cold-Warm		1.0	0.1	
	Wet-Wet		0.1	0.1	0.6
	Dry-Wet		1.2	1.4	
	Dry-Dry		0.1	0.1	0.5
	Wet-Dry		1.1	0.3	
	El Niño-El Niño		0.1	0.1	0.5
	La Niña-El Niño		0.2	0.1	
	La Niña-La Niña		0.1	0.1	0.5
	El Niño-La Niña		0.6	1.2	
Spatial	West-West		0.1	0.1	0.8
	East-West	*	0.1	0.1	
	East-East		0.1	0.1	1.0
	West-East		1.1	1.2	
	North-North		0.1	0.1	0.7
	South-North	*	1.0	0.7	
	South-South		0.1	0.1	0.9
	North-South		0.3	0.3	
Spatio-temporal	Late West-Late West		0.1	0.1	0.8
	Early East-Late West	*	0.9	0.9	
	Early East-Early East		0.1	0.1	1.3
	Late West-Early East		0.4	0.4	
	Early West-Early West		0.1	0.1	0.9
	Late East-Early West	*	0.1	0.1	
	Late East-Late East		0.1	0.1	1.0
	Early West-Late East		0.4	1.0	

Note: Same as Table 3, except for SRT instead of A1SWE.

from building and testing a model in a single basin, it is already well appreciated that the successful calibration and validation of a model in a single basin does not vouch for its spatial generality nor inform us of likely magnitudes of error (e.g., Blöschl et al., 2013; Sivapalan et al., 2003a; Wagener et al., 2001). The community further appreciates that repeating the single basin calibration and validation multiple times in different basins does not demonstrate transferability across space either. Rather, it shows that one can fit a model to each basin and that fit is reliable across time.

Turning our attention to more accepted methods of testing hydrologic model transferability across space, we encounter two remaining possibilities within Klemeš's (1986b) hierarchy of tests: (1) the proxy-basin test (calibrate on basin A and validate on basin B, with A and B in the same region), and (2) the proxy-basin differential split-sample test (records are segregated into, for example, wet and dry periods and the wet period in one basin is used for calibration and the dry period in the other is used for validation). While Klemeš (1986b) suggests that the proxy-basin test is adequate to demonstrate spatial transferability in the context of stable (stationary) conditions, we argue that such a test could select for overly complex models. Recognizing that an event (or year, or similar temporal unit) in two nearby basins could be similarly structured (e.g., in terms of timing and magnitude), we suggest that each event's model error could be similar in the two basins, even if the basins are different (Betterle et al., 2017). This parallels our situation where each site's model error was similar across two time blocks (Figure 8). Fortunately, as we have done here, a hydrologic modeler can test for correlated errors across sampling units (e.g., events or years) to decide whether their validation is potentially biased toward more complex models. Given correlated errors, an alternative validation is the proxy-basin differential split-sample test (aka spatio-temporal NRCV), which splits the sample in time so that one set of events in a given basin is used for calibration and another set of events in a different basin is used for validation. However, as seen in Seibert (2003) and this work (section 4.1.1), validating a highly flexible model using a strong differential can cause the model to fail unless there is at least a little overlap between calibration and validation data ranges.

4.1.3. Degree of Physical Basis

While we billed these models as "empirical", most hydrologic models are, so it is not an entirely useful label. However, there is a contrast between the April 1 SWE model and the SRT model in terms of connection to physical constraints. While the April 1 SWE model can be tied to a clear physical basis, starting with conservation of mass, the SRT model's connection to physical constraints is more tenuous. This difference may contribute to the relative transferability of the two model sets (Table 5 and Figure 9).

April 1 SWE is the water equivalent remaining on the ground much of the way through the typical snow season. It is the difference between the amount of precipitation that fell as snow and the amount that melted in the interim.

$$A1SWE = P - P_{rain} - \text{Melt} \tag{7}$$

Where P_{rain} is the amount of precipitation that falls as rain (November–March), *Melt* is the amount of snowfall that melts prior to April 1, and other variables are as previously defined. P_{rain} represents a fraction of the total precipitation, call it f_rP, and *Melt* represents a fraction of the total snowfall, call it $f_m(P - P_{rain})$; so

$$\begin{aligned} A1SWE &= P - f_rP - f_m(P - f_rP) \\ &= \mathrm{P}(1 - f_r - f_m + f_rf_m) \end{aligned} \tag{8}$$

Where f_r and f_m are constrained on the interval 0–1, and would be expected to be monotonic increasing functions of $\bar{T}$. By conservation of mass, *A1SWE* is constrained to be non-negative but less than or equal to

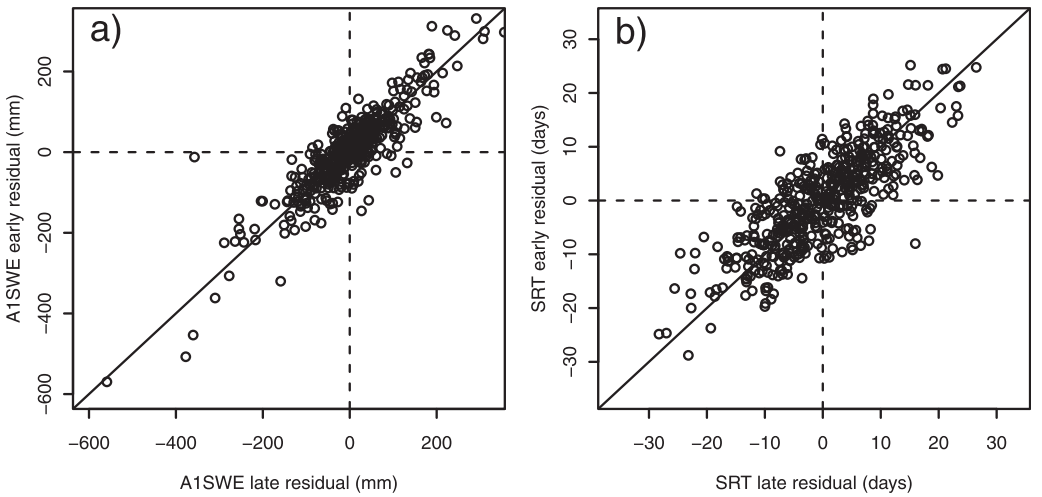

Figure 8. (a) Residuals from the A1SWE spatial analog model calibrated on late years and predicting late-year A1SWE plotted against residuals from the A1SWE model calibrated on early years and predicting early-year A1SWE. (b) Residuals from the SRT spatial analog model calibrated on late years and predicting late-year SRT plotted against residuals from the SRT model calibrated on early years and predicting early-year SRT. Models were calibrated using nn = 0.1 (the most complex formulation). Solid black lines mark the 1:1 line, indicating a perfect match between the residuals. Dashed black lines mark the line of zero prediction error (residual equals zero).

P (within consideration of limits caused by undercatch of snow in precipitation gauges). The term in parentheses on the right of the second equality represents precipitation sensitivity or the increase in SWE per unit addition of precipitation (much like SWE/P, but in the margin):

$$\frac{\partial A1SWE}{\partial P} = 1 - f_r - f_m + f_r f_m \tag{9}$$

This term relates to energy balance components that affect whether precipitation is snow or rain and how much melt occurs. With the previously stated constraints on f_m and f_r, the sensitivity should be on the interval 0–1 and monotonically decrease with increasing temperature, which could be framed as a "statistical-dynamic flux parameterization" (sensu Clark et al., 2017). While temperature can be a dubious substitute for energy balance, we note that temperature tends to be a reasonable predictor of rain-snow partitioning at event time scales (e.g., Dai, 2008), and one can build an argument that if the mean temperature of the winter is colder, the probability distribution of temperatures for the winter, including those during precipitation events is likely colder, and the probability of being below the freezing threshold during a given event is greater. This is essentially the argument used by Woods (2009) to construct a relationship to estimate f_r as a function of mean temperature and precipitation timing, and although that calculation was for an annual scale the same logic can be applied to a seasonal scale. Similarly, temperature is itself a function of energy balance, and more melt tends to happen when temperatures are higher, which is a basis for temperature index melt models. Again, an argument can be built using physical reasoning in a statistical-dynamic framing (rather than necessarily relying on just temperature index approaches) that a warmer mean temperature will tend to have more opportunities and amounts of melt as reflected by an energy balance that keeps a place warmer in winter. Although these arguments can be developed piece by piece to give a hypothetical shape to these relationships (e.g., Woods, 2009), there exists a potential to over-constrain the relationship with our a priori reasoning (sensu Mendoza et al., 2015). We can also be a bit more agnostic about the specifics of the model structure and even what the parameters are (much less their actual value) to explore whether or not this conceptual representation (sensu Gupta & Nearing, 2014) – mass conservation (nominally) with the disposition of that mass a function of mean winter temperature – could be supported by the data if we were to spend the time working out how it should look.

That is the essence of what we have done with the A1SWE models. The model surfaces in Figures 2 and 3 (column 1) can be analyzed for the precipitation sensitivity term on the left side of equation (9), with examples in Figure 11. In Figure 11, we see that while there are clear shapes to the temperature based disposition

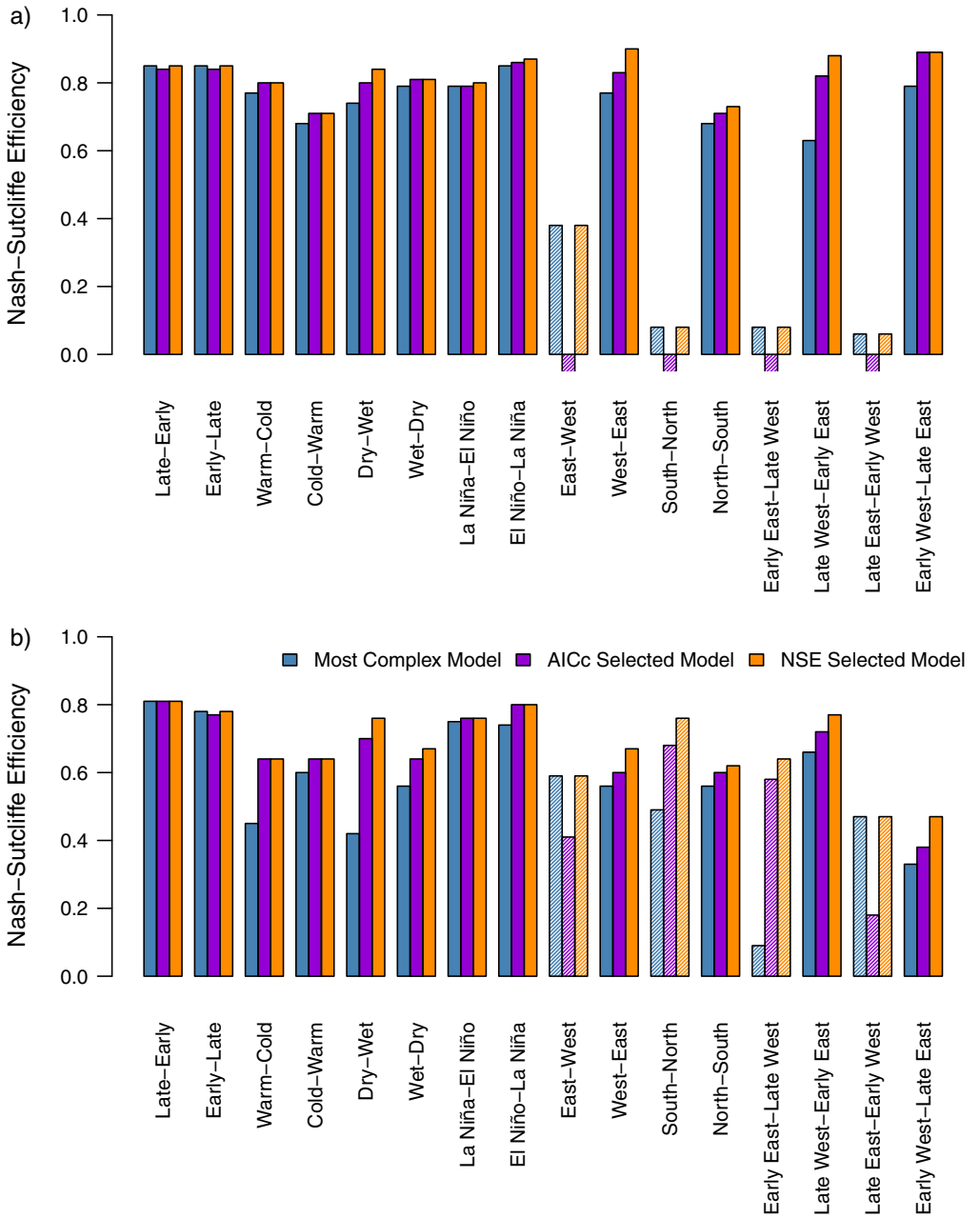

Figure 9. Comparison of NSE values found in validation when using the most complex model (nn = 0.1, in blue), the model suggested by AICc during calibration (in purple), and the model with the best validation NSE (in orange) for (a) April 1 SWE, and (b) SRT. Validation pairs with poor representation in the calibration data are highlighted with cross-hatching.

of precipitation, the curves are somewhat dependent on the calibration data set. These curves and those generated from the NRCV samples are nonlinear, with a plateau of high sensitivity (not far from 1) for temperatures colder than -2°C and a strong decay in how much precipitation is retained as snow on April 1 up to about 4°C, after which little precipitation manifests as April 1 SWE. At the coldest temperatures there is a bit more SWE in April than precipitation (November–March), because some snow may accumulate before November in very cold locations and because of undercatch of snow by precipitation gauges, which limits how strictly we can enforce mass conservation in any model formulation. As seen in Figure 11, there can be differences, however, in the middle range of temperatures that indicate that April 1 SWE is more than a function of $\bar{T}$ and $\bar{P}$ alone.

Table 5
Performance of Most Complex Models Compared to the AICc- and Validation-Selected Models

A1SWE		$NSE_{0.1}$	NSE_{AIC}	NSE_{val}	$KGE_{0.1}$	KGE_{AIC}	KGE_{val}
Late-Early		0.85	0.84	0.85	0.90	0.88	0.90
Early-Late		0.85	0.84	0.85	0.88	0.86	0.88
Warm-Cold		0.77	0.80	0.80	0.84	0.85	0.84
Cold-Warm		0.68	0.71	0.71	0.74	0.74	0.74
Dry-Wet		0.74	0.80	0.84	0.86	0.83	0.89
Wet-Dry		0.79	0.81	0.81	0.82	0.77	0.77
La Niña-El Niño		0.79	0.79	0.80	0.81	0.77	0.79
El Niño-La Niña		0.85	0.86	0.87	0.92	0.93	0.93
East-West	*	0.38	−2.51	0.38	0.68	−0.32	0.68
West-East		0.77	0.83	0.90	0.87	0.86	0.91
South-North	*	0.08	−0.41	0.08	0.30	0.17	0.30
North-South		0.68	0.71	0.73	0.72	0.73	0.76
Early East-Late West	*	0.08	−1.64	0.08	0.42	−0.13	0.42
Late West-Early East		0.63	0.82	0.88	0.81	0.86	0.89
Late East-Early West	*	0.06	−3.40	0.06	0.50	−0.48	0.50
Early West-Late East		0.79	0.89	0.89	0.89	0.91	0.91
SRT		$NSE_{0.1}$	NSE_{AIC}	NSE_{val}	$KGE_{0.1}$	KGE_{AIC}	KGE_{val}
Late-Early		0.81	0.81	0.81	0.82	0.79	0.80
Early-Late		0.78	0.77	0.78	0.89	0.88	0.89
Warm-Cold		0.45	0.64	0.64	0.66	0.74	0.73
Cold-Warm		0.60	0.64	0.64	0.74	0.72	0.72
Dry-Wet		0.42	0.70	0.76	0.63	0.78	0.84
Wet-Dry		0.56	0.64	0.67	0.79	0.81	0.79
La Niña-El Niño		0.75	0.76	0.76	0.83	0.79	0.81
El Niño-La Niña		0.74	0.80	0.80	0.83	0.86	0.87
East-West	*	0.59	0.41	0.59	0.77	0.62	0.77
West-East		0.56	0.60	0.67	0.72	0.76	0.83
South-North	*	0.49	0.68	0.76	0.75	0.80	0.75
North-South		0.56	0.60	0.62	0.66	0.69	0.72
Early East-Late West	*	0.09	0.58	0.64	0.60	0.68	0.69
Late West-Early East		0.66	0.72	0.77	0.83	0.84	0.88
Late East-Early West	*	0.47	0.18	0.47	0.68	0.54	0.68
Early West-Late East		0.33	0.38	0.47	0.63	0.68	0.71

Note: The validation NSE and KGE for the most complex model calibration (subscript 0.1), for the model calibration with the best AICc, and the best validation NSE and validation KGE are shown.

One possible reason for the difference in the ENSO comparison is the timing of precipitation over the course of the season (Woods, 2009). Synchronization of cold temperatures with cold season precipitation, for example, is affected by ENSO phase and region (e.g., McAfee & Wise, 2016), and if precipitation is more focused on shoulder seasons versus mid-winter, the relationship between $\bar{T}$ and precipitation sensitivity would likely change. ENSO phase also affects the joint temperature-precipitation distribution (Lute & Abatzoglou, 2014) and a tendency for wet days to be warmer in one phase than the other would also affect these curves.

Understanding the physical principles involved allows further examination of potential causality for errors found in NRCV validations using plots like Figure 11. Such analyses go beyond the simple error statistics to look at patterns that could elucidate the source of differences. The differences between these curves illustrates the capacity of NRCV tests to highlight model deficiencies that would not be brought to light by LOOCV tests or simple random or sequential temporal split-sample tests. In this way, NRCV tests provide an example of diagnostic model evaluation (de Vos et al., 2010) in which model deficiencies brought to light by the tests are viewed as potential areas for model improvement.

In contrast to the strong physical basis of the A1SWE models, the conceptual basis for the SRT is more intuitive, and the SRT models may be better classified as "empirical" models than physically based. This may well

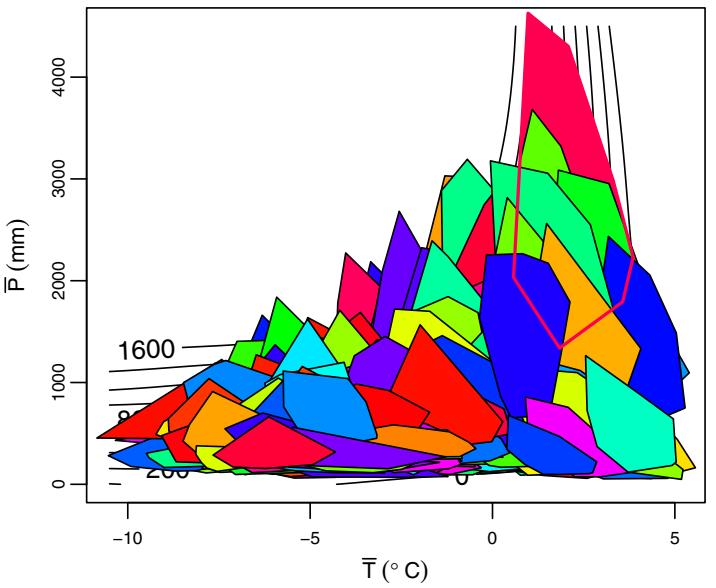

Figure 10. Interannual variability of $\bar{T}$ (°C, x-axis) and $\bar{P}$ (mm, y-axis) at SNOTEL sites (polygons). Each polygon represents a SNOTEL site with the boundary of the polygon encompassing the 21 annual values of $\bar{T}$ and $\bar{P}$ at that site, such that larger polygons indicate greater interannual variability. Polygon colors were randomly assigned. Contour lines behind the polygons are the spatial analog model of April 1 SWE based on all data (1991–2011) and spaced in increments of 200mm. The site with the greatest $\bar{P}$ (red polygon with red lines) is outlined to illustrate the large range of interannual variability at this site.

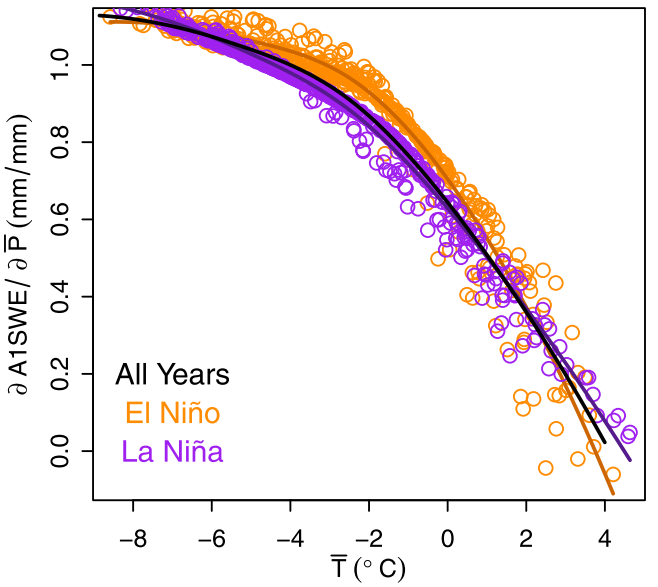

Figure 11. Precipitation sensitivity versus $\bar{T}$ (°C) for spatial analog models built using nn = 0.9 (as selected by AICc in calibration of the El Niño and La Niña models). The black line represents the best fit line for the model calibrated on all years and all sites. The orange (purple) points are the $\bar{T}$ values for El Niño (La Niña) years plotted against the precipitation sensitivities from the model calibrated on El Niño (La Niña) years. The orange (purple) lines are the best fit curves through the orange (purple) points.

be why the validations are poorer for this metric. As has been noted before, the transferability of descriptive empirical models can be poorer than those with a stronger physical basis (e.g., Peel & Blöschl, 2011).

SRT is fundamentally about the temporal juxtaposition of snowfall and snowmelt (see definition sketches in Luce et al. (2014) for example), which relates to temperature distributions. In warm climates, for example, rapid fluctuations around the freezing point can lead to accumulation one week and melt the next. On the other hand in cold climates, where temperatures are persistently below freezing, snowmelt events are rare, and total accumulation depends strongly on seasonal precipitation. Some variability will be tied to this accumulation, as deeper snowpacks take longer to melt, but temperature significantly controls when the melt season begins.

While physical insight can tell us these two variables ($\bar{T}$ and $\bar{P}$) are likely important, and this bears out in validation, the heavier dependence on $\bar{T}$ as a controlling variable, which is usually a proxy for energy balance, means that the physical basis is weaker, and these validations suggest there is potential for improvement by considering other driving variables. SRT is a metric with only recent use, and it is likely that greater insight will be gained as more attempts are made to model it by various means.

4.2. What Level of Model Complexity Was Most Transferable?

Our analysis of optimal model complexity for transferability used a multiple-hypothesis framework (Clark et al., 2011) by simultaneously comparing a range of models of varying complexity. This is in contrast to a common practice in hydrologic modeling of iteratively evaluating model errors and modifying the model (perhaps by adding a new process) to correct for those errors. This practice can be considered a sort of fitting or calibration process and reusing the same dataset to test the revised model would be expected to lead to overfitting to idiosyncrasies. A repetitive validation/diagnosis, fixing/fitting, validation/diagnosis cycle using independent data would also consume data at a prodigious rate, which highlights a strength of multiple hypothesis testing. Within this framework, transfer performance was assessed based on the validation performance alone, rather than performance degradation between calibration and validation as is sometimes done (e.g., Perrin et al., 2001). This is an important distinction since the model with the smallest degradation in performance may not necessarily be the model with the strongest validation (Figure 12).

For the non-temporal tests, including the LOOCV and the spatial and spatiotemporal NRCV tests that did not require significant extrapolation, simpler models transferred best and typically transferred substantially better than the most complex model (Table 5 and Figure 9). These findings add to the list of previous studies, primarily of rainfall-runoff models, that have found that simpler models validate as well or better than more complex models (Chien & Mackay, 2014; Jakeman & Hornberger, 1993; Loague & Freeze, 1985; Perrin et al., 2001; Refsgaard & Knudsen, 1996; Wenger & Olden, 2012). Useful model complexity may be limited by the accuracy of model structures (Perrin et al., 2001) or by data quality (Coxon et al., 2015; Schoups et al., 2008; Westerberg & McMillan, 2015) or resolution (Jakeman & Hornberger, 1993).

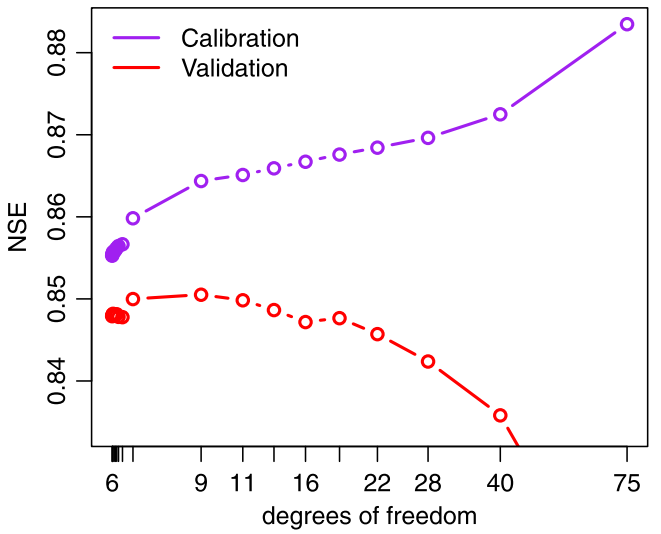

Figure 12. The degrees of freedom (x-axis) associated with the twenty different model complexities (nn = 0.1 to 2.0) are plotted against the calibration (purple line) and validation (red line) NSE values (y-axis) for the spatial analog model of April 1 SWE. Validation used the leave-one-site-out cross-validation test (reference Table 1, line 3). The model with the highest validation NSE value had 9.3 degrees of freedom (nn = 0.9), whereas the model with the least degradation between calibration and validation had 6.0 degrees of freedom (nn = 2.0). The model with the best calibration NSE (nn = 0.1) had the worst validation performance (NSE = 0.81).

Perhaps greater complexity may be warranted in the future as both models and data improve; however, more complex models are typically more difficult to validate due to problems of parameter interpretability (Kirchner, 2006; Wagener et al., 2001) and equifinality (e.g., Beven, 2006). Several have argued, alternatively, that what is needed is "an elegant theory rather than an elaborate, highly parameterized megamodel" (Kirchner, 2006; see also Clark et al., 2016; McDonnell et al., 2007). The models discussed in this paper are empirical and descriptive in nature (sensu Breiman, 2001) but with some theoretical basis in selecting covariates and nominal model structure. They retain a modicum of parsimony, and perhaps elegance, in using only two relatively easy to estimate variables to describe a substantial amount of variability in snowpack and snow retention. Using these models, we demonstrate that even with a large number of sites (n = 497) and a flexible model structure, simpler models perform better.

4.3. Applicability of Results Beyond These Models

Are the behaviors noted above unique to these models, or even this type of algorithmic model? Information-theoretic principles suggest not. Nevertheless it is a good question, and more importantly, it is a testable question following some of the framework applied here. As we note earlier, many hydrologic models are a hybrid of physical reasoning and empirical fitting of unobservable system attributes that are most easily determined from input-output observations. For such models, testing of transferability involves both a calibration and a validation step, and using split-sample and differential split-samples drawn from large long-term data sets is a direct approach for assessment and diagnosis.

Physically based models are an existing standard for predicting snow accumulation and melt and are generally expected to be a useful tool for predicting climate change effects because of their physical basis (e.g., Marty et al., 2017; Musselman et al., 2017; Vionnet et al., 2012). Nevertheless, comparisons across these models in the past using a limited number of sites with detailed weather data, suggest sometimes substantial differences in daily scale predictions (e.g., Etchevers et al., 2004; Rutter et al., 2009). Since much of our interest in climate change is at coarse time scales, there would be value in contrasting several such models when they are applied to climatic averages, as was done in this study. Physically based snow models tend to have a low number of adjustable parameters, even though some use many layers and sub-daily time scales for calculation purposes, which makes the models complicated but not necessarily complex, and likely to perform fairly well. Using a physically based model with fine time steps on a large long-term data set, Parajka et al. (2010) illustrate the value of validation for model assessment and diagnosis. Such a study could be extended to a model selection framework to contrast characteristics such as model complexity across models, similar to the work presented here. Such a rigorous approach to validation would likely benefit complicated models as well by more clearly identifying process representations that are not worth the additional flexibility and overhead.

5. Conclusions

Using a multiple-hypothesis framework, we provide an empirical confirmation of the parsimony principle behind the AICc parameter penalty, validating the information-theoretic approach (e.g., Burnham & Anderson, 2002). While the parsimony principle is well accepted in other modeling cultures (e.g., Burnham & Anderson, 2002; Moreno-Amat et al., 2015), acknowledgement of the value of parsimony is rare in hydrologic modeling literature, where adding another process (and parameter set) is almost *de rigueur* for demonstrating novelty and utility in a model application paper. Similarly, the finding that model transferability can be hampered by model complexity is only rarely seen in hydrologic modeling literature (Jakeman & Hornberger, 1993; Loague & Freeze, 1985; Perrin et al., 2001; Refsgaard & Knudsen, 1996), despite spirited arguments for rigorous validation (Kirchner et al., 1996; Klemeš, 1986b). A survey of the hydrologic literature,

particularly commentaries, leaves an impression that parsimony is a virtue assisting parameter interpretability and reducing the equifinality within a given model (e.g., Jakeman & Hornberger, 1993; Wagener et al., 2001). But equifinality can be a problem even for a simple two parameter physically based infiltration model (e.g., Luce & Cundy, 1994), so this could ultimately be an unsatisfying argument for the benefits of parsimony. According to the information-theoretic approach, and as demonstrated here using empirical models, the better performance of more parsimonious models is a mathematical consequence of the limits of a model and the data with which it is calibrated and confronted.

The model complexities selected by direct validation generally followed those selected by AICc in calibration, implying that AICc is a good guide for model selection (see also Schoups et al., 2008). Differences between direct validation and AICc model selection, however, emphasize the value of validation as a distinct, important diagnostic process. Notably, rigorous evaluation of validation residual patterns can highlight model limitations (see e.g., Kirchner et al., 1996). We further add an example of how incautious sampling (pseudoreplication), inherent in much of the way hydrologists work, can confound decisions regarding optimal model complexity. A focus on process representation in models may yield an implicit bias toward increasing model complexity, so awareness of how our sampling and validation processes work together is important. By paying attention to sampling and thoughtfully applying validation tests in a model selection context, we can objectively and efficiently pursue improvement in hydrologic science.

Acknowledgments
Natural Resources Conservation Service SNOTEL data were obtained from https://wcc.sc.egov.usda.gov/reportGenerator/. TopoWx data were obtained from ftp://mco.cfc.umt.edu/resources/TopoWx-source/auxiliary_data/station_data/. This research was supported in part by an appointment to the U.S. Forest Service Research Participation Program administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and the U.S. Department of Agriculture Forest Service. ORISE is managed by Oak Ridge Associated Universities (ORAU) under DOE contract number DE-SC0014664. All opinions expressed in this paper are the author's and do not necessarily reflect the policies and views of USDA, DOE, or ORAU/ORISE. The insightful comments of three anonymous reviewers greatly improved the quality of this manuscript.

References

Andréassian, V., Perrin, C., Berthet, L., Le Moine, N., Lerat, J., Loumagne, C., . . . Valéry, A. (2009). Crash tests for a standardized evaluation of hydrological models. *Hydrology and Earth System Sciences*, *13*, 1757–1764. https://doi.org/10.5194/hess-13-1757-2009

Betterle, A., Schirmer, M., & Botter, G. (2017). Characterizing the spatial correlation of daily streamflows. *Water Resources Research*, *53*, 1646–1663. https://doi.org/10.1002/2016WR019195

Beven, K. (2001). How far can we go in distributed hydrological modelling? *Hydrology and Earth System Sciences*, *5*, 1–12. https://doi.org/10.5194/hess-5-1-2001

Beven, K. (2006). A manifesto for the equifinality thesis. *Journal of Hydrology*, *320*(11-2), 18–36. https://doi.org/10.1016/j.jhydrol.2005.07.007

Blöschl, G., & Montanari, A. (2010). Climate change impacts-throwing the dice? *Hydrological Processes*, *24*, 374–381. https://doi.org/10.1002/hyp.7574

Blöschl, G., Sivapalan, M., Wagener, T., Viglione, A., & Savenije, H. (Eds.). (2013). *Runoff prediction in ungauged basins: Synthesis across processes, places and scales*. Cambridge, UK: Cambridge University Press.

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, *16*(3), 199–231. https://doi.org/10.1214/ss/1009213726

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). New York, NY: Springer.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel Inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research*, *33*(2), 261–304. https://doi.org/10.1177/0049124104268644

Chien, H., & Mackay, D. S. (2014). How much complexity is needed to simulate watershed streamflow and water quality? A test combining time series and hydrological models. *Hydrological Processes*, *28*, 5624–5636. https://doi.org/10.1002/hyp.10066

Clark, M. P., Bierkens, M. F., Samaniego, L., Woods, R. A., Uijlenhoet, R., Bennett, K. E., . . . Peters-Lidard, C. D. (2017). The evolution of process-based hydrologic models: Historical challenges and the collective quest for physical realism. *Hydrology and Earth System Sciences*, *21*(7), 3427.

Clark, M. P., Kavetski, D., & Fenicia, F. (2011). Pursuing the method of multiple working hypotheses for hydrological modeling. *Water Resources Research*, *47*, W09301. https://doi.org/10.1029/2010WR009827

Clark, M. P., Nijssen, B., Lundquist, J. D., Kavetski, D., Rupp, D. E., Woods, R. A., . . . Brekke, L. D. (2015). A unified approach for process-based hydrologic modeling: 1. Modeling concept, *Water Resources Research*, *51*, 2498–2514. https://doi.org/10.1002/2015WR017200

Clark, M. P., Schaefli, B., Schymanski, S. J., Samaniego, L., Luce, C. H., Jackson, B. M., . . . Istanbulluoglu, E. (2016). Improving the theoretical underpinnings of process-based hydrologic models. *Water Resources Research*, *52*, 2350–2365. https://doi.org/10.1002/2015WR017910

Cooper, M. G., Nolin, A. W., & Safeeq, M. (2016). Testing the recent snow drought as an analog for climate warming sensitivity of the Cascades snowpacks. *Environmental Research Letters*, *11*, https://doi.org/10.1088/1748-9326/11/8/084009

Coron, L., Andréassian, V., Perrin, C., Lerat, J., Vaze, J., Bourqui, M., & Hendrickx, F. (2012). Crash testing hydrological models in contrasted climate conditions: An experiment on 216 Australian catchments. *Water Resources Research*, *48*, W05552. https://doi.org/10.1029/2011WR011721

Coxon, G., Freer, J., Westerberg, I., Wagener, T., Woods, R., & Smith, P. (2015). A novel framework for discharge uncertainty quantification applied to 500 UK gauging stations. *Water Resources Research*, *51*, 5531–5546. https://doi.org/10.1002/2014WR016532

Dai, A. (2008). Temperature and pressure dependence of the rain-snow phase transition over land and ocean. *Geophysical Research Letters*, *35*, L12802. https://doi.org/10.1029/2008GL033295

de Vos, N. J., Rientjes, T. H. M., & Gupta, H. V. (2010). Diagnostic evaluation of conceptual rainfall-runoff models using temporal clustering. *Hydrologic Processes*, *24*, 2840–2850. https://doi.org/10.1002/hyp.7698

Dooge, J. C. I. (1986). Looking for hydrologic laws. *Water Resources Research*, *22*(9S), 46S–58S. https://doi.org/10.1029/WR022i09Sp0046S

Essery, R. L. H., Rutter, N., Pomeroy, J., Baxter, R., Staehli, M., Gustafsson, D., . . . Elder, K. (2009). SnowMIP2: An evaluation of forest snow process simulations. *Bulletin of the American Meteorological Society*, *90*, 1120–1135. https://doi.org/10.1175/2009BAMS2629.1

Etchevers, P. Martin, E., Brown, R., Fierz, C., Lejeune, Y., Bazile, E., . . . Yamazaki, T. (2004). Validation of the energy budget of an alpine snowpack simulated by several snow models (SnowMIP project). *Annals of Glaciology*, *38*, 150–158. https://doi.org/10.3189/172756404781814825

Franchini, M., & Pacciani, M. (1991). Comparative analysis of several conceptual rainfall-runoff models. *Journal of Hydrology*, *122*(1–4), 161–219. https://doi.org/10.1016/0022-1694(91)90178-K

Gupta, H. V., Clark, M. P., Vrugt, J. A., Abramowitz, G., & Ye, M. (2012). Towards a comprehensive assessment of model structural adequacy. *Water Resources Research*, *48*, W08301. https://doi.org/10.1029/2011WR011044

Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modeling. *Journal of Hydrology*, *377*, 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003

Gupta, H. V., & Nearing, G. S. (2014). Debates—The future of hydrological sciences: A (common) path forward? Using models and data to learn: A systems theoretic perspective on the future of hydrological science. *Water Resources Research*, *50*, 5351–5359. https://doi.org/10.1002/2013WR015096

Gupta, H. V., Perrin, C., Bloschl, G., Montanari, A., Kumar, R., Clark, M., & Andreassian, V. (2014). Large-sample hydrology: A need to balance depth with breadth. *Hydrology and Earth System Sciences*, *18*, 463–477. https://doi.org/10.5194/hess-18-463-2014

Gupta, H. V., Wagener, T., & Liu, Y. (2008). Reconciling theory with observations: Elements of a diagnostic approach to model evaluation. *Hydrological Processes*, *22*, 3802–3813. https://doi.org/10.1002/hyp.6989

Hay, L. E., & Clark, M. (2003). Use of statistically and dynamically downscaled atmospheric model output for hydrologic simulations in three mountainous basins in the western United States. *Journal of Hydrology*, *282*(1–4), 56–75. https://doi.org/10.1016/S0022-1694(03)00252-X

Hrachowitz, M., & Clark, M. (2017). HESS Opinions: The complementary merits of competing modelling philosophies in hydrology. *Hydrology and Earth System Sciences*, *21*, 3953–3973. https://doi.org/10.5194/hess-2017-36

Hrachowitz, M., Savenije, H. H. G., Blöschl, G., McDonnell, J. J., Sivapalan, M., Pomeroy, J. W., . . . Cudennec, C. (2013). A decade of predictions in ungauged basins (PUB)—A review. *Hydrological Sciences Journal*, *58*(6), 1198–1255. https://doi.org/10.1080/02626667.2013.803183

Hurlbert, S. H. (1984). Pseudoreplication and the design of ecological field experiments. *Ecological Monographs*, *54*(2), 187–211. https://doi.org/10.2307/1942661

Hurvich, C. M., & Tsai, C.-L. (1989). Regression and time series model selection in small sample. *Biometrika*, *76*(2), 99–104. https://doi.org/https://doi.org/10.1093/biomet/76.2.297

Jakeman, A. J., & Hornberger, G. M. (1993). How much complexity is warranted in a rainfall-runoff model? *Water Resources Research*, *29*(8), 2637–2649. https://doi.org/10.1029/93WR00877

Kirchner, J. W. (2006). Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology. *Water Resources Research*, *42*, W03S04. https://doi.org/10.1029/2005WR004362

Kirchner, J. W., Hooper, R. P., Kendall, C., Neal, C., & Leavesley, G. (1996). Testing and validating environmental models. *The Science of the Total Environment*, *183*, 33–47. https://doi.org/10.1016/0048-9697(95)04971-1

Klemeš, V. (1986a). Dilettantism in hydrology: Transition or destiny? *Water Resources Research*, *22*(9), 177S–188S. https://doi.org/10.1029/WR022i09Sp0177S

Klemeš, V. (1986b). Operational testing of hydrological simulation models. *Hydrological Sciences Journal*, *31*(1), 13–24. https://doi.org/10.1080/02626668609491024

Li, C. Z., Zhang, L., Wang, H., Zhang, Y. Q., Yu, F. L., & Yan, D. H. (2012). The transferability of hydrological models under nonstationary climatic conditions. *Hydrology and Earth System Sciences*, *16*, 1239–1254. https://doi.org/10.5194/hess-16-1239-2012

Liang, X., Lettenmaier, D. P., Wood, E. F., & Burges, S. J. (1994). A simple hydrologically based model of land surface water and energy fluxes for general circulation models. *Journal of Geophysical Research*, *99*(D7), 14415–14428. https://doi.org/10.1029/94JD00483

Lima, C. H. R., & Lall, U. (2010). Spatial scaling in a changing climate: A hierarchical Bayesian model for non-stationary multi-site annual maximum and monthly streamflow. *Journal of Hydrology*, *383*(3–4), 307–318. https://doi.org/10.1016/j.jhydrol.2009.12.045

Loader, C. (2013). Locfit: Local regression, likelihood and density estimation, in R package version 1.5–9.1. Retrieved from http://CRAN.R-project.org/package=locfit

Loader, C. R. (1999). *Local regression and likelihood*. New York, NY: Springer.

Loague, K. M., & Freeze, R. A. (1985). A comparison of rainfall-runoff modeling techniques on small upland catchments. *Water Resources Research*, *21*(2), 229–248. https://doi.org/10.1029/WR021i002p00229

Luce, C. H., & Cundy, T. W. (1994). Parameter identification for a runoff model for forest roads. *Water Resources Research*, *30*(4), 1057–1069.

Luce, C. H., Lopez-Burgos, V., & Holden, Z. (2014). Sensitivity of snowpack storage to precipitation and temperature using spatial and temporal analog models. *Water Resources Research*, *50*, 9447–9462. https://doi.org/10.1002/2013WR014844

Lute, A. C., & Abatzoglou, J. T. (2014). Role of extreme snowfall events in interannual variability of snowfall accumulation in the western United States. *Water Resources Research*, *50*, 2874–2888. https://doi.org/10.1002/2013WR014465

Marty, C., Schlogl, S., Bavay, M., & Lehning, M. (2017). How much can we save? Impact of different emission scenarios on future snow cover in the Alps. *Cryosphere*, *11*(1), 517–529. https://doi.org/10.5194/tc-11-517-2017

McAfee, S. A., & Wise, E. K. (2016). Intra-seasonal and inter-decadal variability in ENSO impacts on the Pacific Northwest. *International Journal of Climatology*, *36*, 508–516. https://doi.org/10.1002/joc.4351

McDonnell, J. J. *et al.* (2007). Moving beyond heterogeneity and process complexity: A new vision for watershed hydrology. *Water Resources Research*, *43*, W07301. https://doi.org/10.1029/2006WR005467

McMillan, H., Montanari, A., Cudennec, C., Savenije, H., Kreibich, H., Krueger, T., . . . Aksoy, H. (2016). Panta Rhei 2013–2015: Global perspectives on hydrology, society and change. *Hydrological Sciences Journal*, *61*(7), 1174–1191. https://doi.org/10.1080/02626667.2016.1159308

Mendoza, P. A., Clark, M. P., Barlage, M., Rajagopalan, B., Samaniego, L., Abramowitz, G., & Gupta, H. (2015). Are we unnecessarily constraining the agility of complex process-based models? *Water Resources Research*, *51*, 716–728. https://doi.org/10.1002/2014WR015820

Menne, M. J., & Williams, C. N. (2009). Homogenization of temperature series via pairwise comparisons. *Journal of Climate*, *22*(7), 1700–1717. https://doi.org/10.1175/2008JCLI2263.1

Merz, R., Parajka, J., & Bloschl, G. (2011). Time stability of catchment model parameters: Implications for climate impact analyses. *Water Resources Research*, *47*, W02531. https://doi.org/10.1029/2010WR009505

Milly, P. C. D., Betancourt, J. L., Falkenmark, M., Hirsch, R. M., Kundzewicz, Z. W., Lettenmaier, D. P., & Stouffer, R. J. (2008). Climate change: Stationarity is dead: Whither water management? *Science*, *319*, 573–574. https://doi.org/10.1126/science.1151915

Montanari, A., Young, G., Savenije, H., Hughes, D., Wagener, T., Ren, L., . . . Grimaldi, S. (2013). Panta Rhei-everything flows'': Change in hydrology and society-the IAHS scientific decade 2013–2022. *Hydrological Sciences Journal*, *58*(6), 1256–1275. https://doi.org/10.1080/02626667.2013.809088

Moreno-Amat, E., Mateo, R. G., Nieto-Lugilde, D., Morueta-Holme, N., Svenning, J.-C., & Garcia-Amorena, I. (2015). Impact of model complexity on cross-temporal transferability in Maxent species distribution models: An assessment using paleobotanical data. *Ecological Modelling*, *312*, 308–317. https://doi.org/10.1016/j.ecolmodel.2015.05.035

Mote, P. W. (2006). Climate-driven variability and trends in mountain snowpack in western North America*. *Journal of Climate*, *19*, 6209–6220. https://doi.org/10.1175/JCLI3971.1

Mote, P. W., Rupp, D. E., Li, S., Sharp, D. J., Otto, F., Uhe, P. F., . . . Allen, M. F. (2016). Perspectives on the causes of exceptionally low 2015 snowpack in the western United States. *Geophysical Research Letters*, *43*, 10980–10988. https://doi.org/10.1002/2016GL069965

Musselman, K. N., Clark, M. P., Liu, C. H., Ikeda, K., & Rasmussen, R. (2017). Slower snowmelt in a warmer world. *Nature Climate Change*, *7*(3), 214–219. https://doi.org/10.1038/Nclimate3225

Oyler, J. W., Ballantyne, A., Jencso, K., Sweet, M., & Running, S. W. (2014). Creating a topoclimatic daily air temperature dataset for the conterminous United States using homogenized station data and remotely sensed land skin temperature. *International Journal of Climatology*, *35*, 2258–2279. https://doi.org/10.1002/joc.4127

Oyler, J. W., Dobrowski, S. Z., Ballantyen, A. P., Klene, A. E., & Running, S. W. (2015). Artificial amplification of warming trends across the mountains of the western United States. *Geophysical Research Letters*, *42*, 153–161. https://doi.org/10.1002/2014GL062803

Parajka, J., Dadson, S., Lafon, T., & Essery, R. (2010). Evaluation of snow cover and depth simulated by a land surface model using detailed regional snow observations from Austria. *Journal of Geophysical Research*, *115*, D24117. https://doi.org/10.1029/2010JD014086

Peel, M. C., & Blöschl, G. (2011). Hydrological modelling in a changing world. *Progress in Physical Geography, 32*(2), 249–261. https://doi.org/10.1177/0309133311402550

Perrin, C., Michel, C., & Andréassian, V. (2001). Does a large number of parameters enhance model performance? Comparative assessment of common catchment model structures on 429 catchments. *Journal of Hydrology*, *242*, 274–301. https://doi.org/10.1016/S0022-1694(00)00393-0

Refsgaard, J. C., & Knudsen, J. (1996). Operational validation and intercomparison of different types of hydrological models. *Water Resources Research*, *32*(7), 2189–2202. https://doi.org/10.1029/96WR00896

Refsgaard, J. C., Madsen, H., Andréassian, V., Arnbjerg-Nielsen, K., Davidson, T. A., Drews, M., . . . Christensen, J. H. (2014). A framework for testing and ability of models to project climate change and its impacts. *Climatic Change*, *122*, 271–282. https://doi.org/10.1007/s10584-013-0990-2

Rutter, N., Essery, R., Pomeroy, J., Altimir, N., Andreadis, K., Baker, I., . . . Yamazaki, T. (2009). Evaluation of forest snow processes models (SnowMIP2). *Journal of Geophysical Research*, *114*, D06111. https://doi.org/10.1029/2008JD011063

Schoups, G., van de Giesen, N. C., & Savenije, H. H. G. (2008). Model complexity control for hydrologic prediction. *Water Resources Research*, *44*, W00B03. https://doi.org/10.1029/2008WR006836

Seibert, J. (2003). Reliability of model predictions outside calibration conditions. *Nordic Hydrology*, *34*(5), 477–492.

Singh, R., Wagener, T., van Werkhoven, K., Mann, M. E., & Crane, R. (2011). A trading-space-for-time approach to probabilistic continuous streamflow predictions in a changing climate- accounting for changing watershed behavior. *Hydrology and Earth System Sciences*, *15*, 3591–3603. https://doi.org/10.5194/hess-15-3591-2011

Sivapalan, M. (2009). The secret to 'doing better hydrological science': Change the question! *Hydrological Processes*, *23*, 1391–1396. https://doi.org/10.1002/hyp.7242

Sivapalan, M., Bloschl, G., Zhang, L., & Vertessy, R. (2003b). Downward approach to hydrological prediction. *Hydrological Processes*, *17*(11), 2101–2111. https://doi.org/10.1002/hyp.1425

Sivapalan, M. Takeuchi, K., Franks, S. W., Gupta, V. K., Karambiri, H., Lakshmi, V., Liang, X., . . . Zehe, E. (2003a). IAHS Decade on Predictions in Ungauged Basins (PUB), 2003–2012: Shaping an exciting future for the hydrological sciences. *Hydrological Sciences Journal*, *48*(6), 857–880. https://doi.org/10.1623/hysj.48.6.857.51421

Sproles, E. A., Roth, T. R., & Nolin, A. W. (2017). Future snow? A spatial-probabilistic assessment of the extraordinarily low snowpacks of 2014 and 2015 in the Oregon Cascades. *The Cryosphere*, *11*(1), 331–341. https://doi.org/10.5194/tc-11-331-2017

Thirel, G. Andréassian, V., Perrin, C., Audouy, J.-N., Berthet, L., Edwards, P., . . . Vaze, J. (2015). Hydrology under change: An evaluation protocol to investigate how hydrological models deal with changing catchments. *Hydrological Sciences Journal*, *60*(7–8), 2150–3435. https://doi.org/10.1080/02626667.2014.967248

Vionnet, V., Brun, E., Morin, S., Boone, A., Faroux, S., Le Moigne, P., Martin, E., & Willemet, J. M. (2012). The detailed snowpack scheme Crocus and its implementation in SURFEX v7.2. *Geoscientific Model Development*, *5*(3), 773–791. https://doi.org/10.5194/gmd-5-773-2012

Wagener, T., Boyle, D. P., Lees, M. J., Wheater, H. S., Gupta, H. V., & Sorooshian, S. (2001). A framework for development and application of hydrological models. *Hydrology and Earth System Sciences*, *5*(1), 13–26. https://doi.org/10.5194/hess-5-13-2001

Wagener, T., & Gupta, H. V. (2005). Model identification for hydrological forecasting under uncertainty. *Stochastic Environmental Research and Risk Assessment*, *19*(6), 378–387. https://doi.org/10.1007/s00477-005-0006-5

Wenger, S. J., & Olden, J. (2012). Assessing transferability of ecological models: An underappreciated aspect of statistical validation. *Methods in Ecology and Evolution*, *3*, 260–267. https://doi.org/10.1111/j.2041-210X.2011.00170.x

Wenger, S. J. Som, N. A., Dauwalter, D. C., Isaak, D. J., Neville, H. M., Luce, C. H., . . . Rieman, B. E. (2013). Probabilistic accounting of uncertainty in forecasts of species distributions under climate change. *Global Change Biology*, *19*, 3343–3354. https://doi.org/10.1111/gcb.12294

Westerberg, I., & McMillan, H. (2015). Uncertainty in hydrological signatures. *Hydrology and Earth System Sciences*, *19*(9), 3951–3968.

Wilby, R. L., & Wigley, T. (1997). Downscaling general circulation model output: A review of methods and limitations. *Progress in Physical Geography*, *21*(4), 530–548.

Wolter, K., & Timlin, M. S. (1993). Monitoring ENSO in COADS with a seasonally adjusted principal component index. In *Proceedings of the 17th climate diagnostics workshop* (pp. 52–57). Norman, OK: NOAA/NMC/CAC, NSSL, Oklahoma Climatological Survey, CIMMS and the School of Meteorology, University of Oklahoma.

Woods, R. A. (2009). Analytical model of seasonal climate impacts on snow hydrology: Continuous snowpacks. *Advances in Water Resources*, *32*(10), 1465–1481. https://doi.org/10.1016/j.advwatres.2009.06.011

Young, P. (2003). Top-down and data-based mechanistic modelling of rainfall-flow dynamics at the catchment scale. *Hydrological Processes*, *17*, 2195–2217. https://doi.org/10.1002/hyp.1328