Bioclimatic modelling using Gaussian mixture distributions and multiscale segmentation

Daniel G. Gavin* and Feng Sheng Hu

ABSTRACT

Aim To introduce Gaussian mixture distributions and sequential maximum a posteriori image segmentation (GM-SMAP) as a model that predicts species ranges from mapped climatic variables, and to compare its predictive capacity with two commonly used bioclimatic models: regression tree analysis (RTA) and smoothed response surfaces (SRS).

Location North-west North America.

Methods We compared models for their ability to predict the distributional range of western hemlock (*Tsuga heterophylla*). We calculated and projected nine climatic and water-balance variables to a 2-km grid up to 140 km from the *T. heterophylla* range. Models were trained using the five variables selected by RTA, as well as subsets of three variables. Goodness of fit was assessed using models trained and tested on the entire study area. Predictive capacity was assessed using 100 cross-validation tests, each trained on a randomly sampled 1% of the study area and tested on the complement of the study area.

Results Models using all five variables were significantly better than three-variable models. Model fit was greatest for SRS. GM-SMAP misclassified slightly more area and RTA misclassified almost twice the area compared to SRS. However, cross-validation showed that the predictive capacity was clearly greatest for GM-SMAP and lowest for SRS, indicating that GM-SMAP makes more accurate predictions from sparse data.

Main conclusions GM distributions prevent overfitting using an information-theoretic approach, and the SMAP algorithm minimizes the spatial extent of the largest misclassified area using a multiscale method. These properties, useful for image classification, also aid their strong predictive capacity as a bioclimatic model. SRS overfit the data, lowering its predictive capacity, and RTA failed to capture details of interactions among variables, yielding a poor fit. These results demonstrate the strong potential of GM-SMAP as a bioclimatic model.

Keywords Actual evapotranspiration, bioclimatic envelope models, ecological niche, model selection, regression tree analysis, response surfaces, *Tsuga heterophylla*, western hemlock.

INTRODUCTION

Models of the relationship between the observed range of a species and mapped climatic and/or climate-derived variables (ecological niche models or bioclimatic models) are widely used to examine the climatic controls of species range limits and to predict species ranges under future climatic conditions (Scott et al., 2003; Thomas et al., 2004). Although criticized for their simplistic assumption of a strong association between regional climate and species distributions (Loehle & LeBlanc, 1996; Woodward & Beerling, 1997; Hampe, 2004), these models remain useful for understanding the potential for climatic control of range limits and may identify instances where non-climatic factors are important (Pearson & Dawson, 2003; Huntley et al., 2004).
There are many functional forms of bioclimatic models, and the choice of one model over another may have a large effect on the conclusions of a study (Franklin, 1995; Guisan & Zimmermann, 2000; Austin, 2002). However, few studies have compared the strengths and weaknesses of different bioclimatic models (Segurado & Araújo, 2004).

Our overall goal in this study is to demonstrate the performance of an image classification method as a bioclimatic model. This method, introduced by Bouman and Shapiro (1994) and Bouman (1998), first fits a Gaussian mixture distribution to training data and then 'segments' maps or images into regions using sequential maximum a posteriori image segmentation (hereafter termed GM-SMAP). The GM-SMAP model was developed primarily for segmenting multiband satellite images into regions with a certain behaviour or spectral signature. Modelling land cover classes using satellite imagery bands is statistically analogous to modelling a species range using mapped climatic variables, though species distribution mapping requires only two classes (species presence and absence). In fact, many models have been used recently for both image segmentation and bioclimatic modelling, including artificial neural networks (Pearson et al., 2002), regression trees (Iverson & Prasad, 2001), generalized additive models (Zaniewski et al., 2002) and genetic algorithms (Oberhauser & Peterson, 2003). Because in remote sensing studies, GM-SMAP outperforms several other methods of image segmentation (McCaulley & Engel, 1995; Michelson et al., 2000), we expected that it would be promising as a bioclimatic model.

A key consideration for choosing a bioclimatic model and fitting it to training data is the degree to which the model fits the data structure. If predictor variables have a direct functional control on species presence, then simple thresholds on individual variables could predict species presence. However, it is likely that individual variables alone cannot adequately describe the species–environment relationship, but that variable interactions may approximate an unknown variable with a more direct functional control on the species. Thus, the efficacy of a model depends on how closely it fits the structural complexity of the training data set. Overfitting the training data will result in low predictive capacity, i.e., the ability to predict in regions other than where the model has been trained, and underfitting will not capture meaningful species–environment correlations (Guisan & Zimmerman, 2000). Good fit and predictive capacity are challenging to achieve when using large data sets that capture many details of the species–environment relationship, such as provided by high-resolution maps of large areas. The GM-SMAP model optimizes the fit by using an information-theoretic approach and by mapping predictions in the context of neighbouring grid points. These characteristics are addressed in detail below.

To evaluate the GM-SMAP model, we compare it to two commonly used bioclimatic models: regression tree analysis (RTA) and smoothed response surfaces (SRS). One way to compare these models would be to examine their ability to detect structures in simulated data sets. A recent attempt using this approach to rank 10 multivariate nonparametric regression models found that it was very difficult to prescribe models for specific data structures (Banks et al., 2003). That study concluded with the recommendation to compare models using a portion of the training data and examining the fit to the 'holdouts' (i.e. cross-validation) rather than use arguments based on prior knowledge of the form of the model and the structure of the data. In this study, we follow this recommendation by comparing the capacity of the three models to predict the range of a single species: western hemlock (Tsuga heterophylla). By focusing on a single species, we cannot explore the generality of GM-SMAP in other settings. However, T. heterophylla serves as a good test case because it is similar to several tree species in the size of its distribution and its juxtaposition with various climatic regions.

We evaluated the three methods for bioclimatic modelling in two steps. First, we trained the models on three sets of bioclimatic variables to evaluate the effects of different variable choices. Model accuracy was assessed by examining the goodness of fit of models trained on the entire study area (resubstitution). Second, we compared the predictive capacity of the models using a cross-validation method where models were trained on random subsamples of the study area and tested on the complement of the study area. The cross-validation analysis is a robust measure of a model's performance because it demonstrates the capacity to make predictions on new data sets. In contrast, the resubstitution test cannot assess if the model is overfitting the data (Fielding & Bell, 1997).

The GM-SMAP segmentation model

Modelling species ranges using the GM-SMAP model is a multistep process outlined in Fig. 1. Here we describe these algorithms in non-mathematical intuitive terms; complete formulations are in Bouman and Shapiro (1994) and Bouman (1998). The first step is to fit the training data for two classes (species presence and absence) with a GM distribution using the program cluster.
version 3.5.4 (Bouman, 1998). A GM is a probabilistic model composed of a number of subclasses, each described by a multivariate Gaussian distribution (Fig. 2a). Each multivariate Gaussian distribution is defined by a small number of parameters (the mean and variance of each variable, the covariances between each pair of variables, and a weighting based on the proportion of data described by the subclass). While a multivariate Gaussian distribution defines each subclass, the combined mixture of distributions adapts to nonlinear patterns in the data and thus the GM for each class does not resemble a Gaussian distribution.

The number of subclasses in the GM can be specified either a priori or estimated directly from the data. The latter approach is appealing because the tightness of the fit of the GM can be determined objectively to avoid overfitting or underfitting as described below.

The GM distribution is fit by first initializing a large number of subclasses with the means, covariances and ‘weightings’ of each subclass. This ‘seed’ GM is made of randomly selected means and identical variance-covariance matrices based on the entire data set. Subclasses are then modified using the iterative expectation maximization (EM) algorithm of Dempster et al. (1977). This algorithm (1) estimates the probability of each observation in the training data belonging to each subclass (using the GM parameters) then (2) re-computes the maximum-likelihood estimate of the GM parameters using the probabilities from step 1. These two steps are repeated until the GM parameters converge on a final maximum-likelihood estimate.

The steps described above computes a GM for a predetermined number of subclasses, but they do not address how to determine the ‘best’ number of subclasses in a GM. Adding more subclasses always increases the fit of the GM, but too many subclasses would record fine-scale structures in the data resulting in overfitting. The cluster program guards against overfitting by using the minimum description length (MDL) estimator (Grünwald, 2004), a selection method that weighs complexity in the model against complexity in the data. MDL is calculated by penalizing the maximum likelihood of the model with increases in the number of model parameters and the sample size (Bouman, 1998). This procedure prevents models from adding more subclasses when the additional complexity yields diminishing returns in increased fit. As applied here, MDL is used to compare the fit of GM distributions composed of different numbers of subclasses. The number of subclasses is reduced from the initial large number by combining the two nearest subclasses and running the EM algorithm on the new set of subclasses. This agglomeration of subclasses is repeated until only one subclass remains, and the number of subclasses retained in the final GM is the one with the smallest MDL.

Predictions from the GM distribution are projected onto mapped variables using the sequential maximum a posteriori (SMAP) algorithm, available in GRASS GIS software (Bouman & Shapiro, 1994). This Bayesian approach is a type of Markov chain model applied across resolutions or scales. The SMAP algorithm segments grid-point class labels (i.e. species presence or absence) at multiple resolutions using (1) the class label at the previous coarser resolution, (2) the GM distribution, and (3) a cost function. The cost function, a measure of the probability of a coarse-resolution label changing at finer resolutions, is calibrated by a fine-to-coarse procedure in which class labels at each resolution are compared to the data at finer resolutions. The cost function puts exponentially greater weight on errors at coarse resolutions because those errors correspond to larger areas. Both the fine-to-coarse cost function calibration and the coarse-to-fine segmentation are applied to the gridded variables using a structure that defines the neighbourhood relationships across resolutions (Fig. 3). At fine resolutions, this structure is a quadtree, and each grid point depends on one coarse-resolution neighbour (Fig. 3a). If applied to coarse resolutions, the quadtree structure would...
result in unrealistically blocky regions. Thus, at coarse resolutions, the quadtree structure is replaced by an augmented pyramidal structure in which each grid point is dependent on three coarse-resolution neighbours (Fig. 3b). This structure results in smooth region boundaries during the coarse-to-fine segmentation. Details of these calculations are presented in Bouman and Shapiro (1994) and software for \textsc{cluster} and \textsc{smap} is freely available (http://www.ece.purdue.edu/~bouman).

The net effect of the SMAP algorithm is to make a more spatially contiguous classification than would be achieved if each grid point were classified independently using a maximum-likelihood method. The cost function adapts to the scale of heterogeneity (patchiness) in the data, and thus does not overly smooth classifications. Therefore, SMAP maximizes the accuracy of predictions by minimizing the size of the largest misclassified region and creating predictions with the level of spatial autocorrelation that occurs in species range maps.

\section*{METHODS}

\subsection*{Zonal ecosystems and bioclimatic variables}

We trained models on zonal ecosystems with \textit{T. heterophylla} as a dominant species. Zonal ecosystems, sometimes termed biogeoclimatic zones, are defined as areas with relatively homogeneous late successional vegetation in the absence of local controls, such as atypical soil or drainage (Meidinger & Pojar, 1991). Models trained on zonal ecosystem capture regional-scale bioclimatic variations that define the zonal ecosystem. The perimeter of all known occurrences of a species based on actual (plot-based) observations is less appropriate because it encompasses diffuse range limits where microclimates differ from the regional climate and where biotic interactions may be strong. Zonal ecosystems are mapped by interpolating between field plots using locally derived elevation and aspect rules, and thus can capture elevationally stratified forest zones at scales of < 500 m (Meidinger & Pojar, 1991). To create a 2-km resolution \textit{T. heterophylla} map, we merged four widely used ecosystem classifications developed by various government agencies:

1. Broad Ecosystem Inventory by the British Columbia Ministry of Sustainable Resource Management, covering British Columbia, southeast Alaska, Washington and Idaho (85.2% of the \textit{T. heterophylla} zone; http://www.gov.bc.ca/ecology/bei/shiningmnts.html)
2. Ecoregions of Oregon by the United States Geological Survey (13.7%; http://www.gis.state.or.us/data/alphabet.html)
3. Alaska Statewide Vegetation by the United States Geological Survey (0.7%; http://agdc.usgs.gov/data/projects/hlct/hlct.html)
4. California Vegetation Maps by the United States Forest Service (0.4%; http://gis.ca.gov/catalog/BrowseRecord.epl?id=708).

Climatic and water-balance variables were developed from the monthly normals available from the PRISM climate mapping project (Daly et al., 1994). These peer-reviewed maps were created by interpolating weather station data with elevation as a covariate and with algorithms to estimate rain shadow effects, thermal inversions and coastal climates. We initially examined nine potential bioclimatic variables (Table 1), focusing on summer moisture stress, which is a known physiological limit on the \textit{T. heterophylla} range (Waring & Franklin, 1979; Lassoie et al., 1986). Potential annual evapotranspiration (PET) and actual annual evapotranspiration (AET) were estimated using the water-balance model presented in Willmott et al. (1985). All climatic variables and the water-balance model were calculated for each 2-km grid point. The maximum correlation coefficient was 0.66 between any pair of variables. We did not include the annual climatic moisture deficit (PET-AET; Stephenson, 1998) because it was highly correlated ($r = 0.98$) with the widely used AET/PET moisture index (Huntley et al., 1995).

\subsection*{RTA and SRS bioclimatic models}

We compared model predictions from the GM-SMAP model (described above) to regression tree analysis (RTA) and smoothed response surfaces (SRS). The RTA model classifies grid points using a branching decision tree (identical to classification and
regression trees, CART; Iverson & Prasad, 2001). It was run using a software integrated with GRASS GIS (Therneau & Atkinson, 1997) following the methods of Breiman et al. (1984). Each split in the decision tree represents the threshold of the variable from a large list of candidate variables (Table 1) that most accurately classifies grid points. This results in a set of thresholds for each variable that are dependent on values of other variables (Fig. 2b). For this study, the regression tree was grown until further branching did not improve the overall classification better than 0.25%, and then the tree was pruned back by using a cross-validation method to remove spurious branching (Breiman et al., 1984).

The SRS model develops a response surface based on the relative frequency of grid points with species present vs. absent in climate space (Huntley et al., 1989). Our use of SRS (using a custom computer program) is similar to locally weighted regression (LOESS) and nonparametric multiplicative regression presented in McCune et al. (2003). Rather than developing a response surface on an arbitrary lattice within climate space and making predictions from values in this lattice, we calculated the exact response surface value for each grid point to be predicted. For each grid point, the number of additional grid points within a specific climatic window was calculated. The window width of each variable was approximately 20% of the central 95% range within the T. heterophylla zone (Table 1). Each point falling within this window was weighted using the tricube filter (Huntley et al., 1989), so that grid points with a very similar climate were weighted more heavily than points closer to the limits of the window (Fig. 2c). The same weighted sum was calculated both for the points within the T. heterophylla zone and for all points within and outside the T. heterophylla zone. The response surface value was calculated as the ratio of these two values, and ranged between 0 (climate at the grid point is unlike any in the T. heterophylla zone) to 1 (climate at the grid point is unlike any not in the T. heterophylla zone). Thresholds for determining T. heterophylla presence were set where the underprediction (omission) error rate was at least as low as for other models and lower than the overprediction (commission) error rate.

### Modelling strategy

As GM-SMAP and SRS do not automatically determine which variables to include in the model, we compared model fits using three sets of climatic variables. The first set of variables is widely used in other modelling studies (e.g. Huntley et al., 1995; Shafer et al., 2001): AET/PET (actual/potential annual evapotranspiration),

<table>
<thead>
<tr>
<th>Variable</th>
<th>Code</th>
<th>2.5th–97.5th percentiles (T. heterophylla zone and total region)*</th>
<th>Window width for response surfaces†</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean temperature of the coldest month (°C)</td>
<td>MTCO</td>
<td>−10.8–5.5</td>
<td>4.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>−20.2–4.5</td>
<td></td>
</tr>
<tr>
<td>Growing degree-days on a 5° base‡</td>
<td>GDD5</td>
<td>515–2425</td>
<td>400</td>
</tr>
<tr>
<td></td>
<td></td>
<td>107–2511</td>
<td></td>
</tr>
<tr>
<td>Actual annual evapotranspiration (mm)§</td>
<td>AET</td>
<td>384–608</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td></td>
<td>238–580</td>
<td></td>
</tr>
<tr>
<td>Moisture index</td>
<td>AET/PET</td>
<td>0.699–0.995</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.398–0.995</td>
<td></td>
</tr>
<tr>
<td>Annual precipitation (mm)</td>
<td>P</td>
<td>650–4930</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>225–5073</td>
<td></td>
</tr>
<tr>
<td>Effective precipitation (mm)</td>
<td>P-PET</td>
<td>145–4418</td>
<td>800</td>
</tr>
<tr>
<td></td>
<td></td>
<td>−320–4621</td>
<td></td>
</tr>
<tr>
<td>July precipitation (mm)</td>
<td>P07</td>
<td>11–20</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>4–250</td>
<td></td>
</tr>
<tr>
<td>Snow water equivalent (mm)§</td>
<td>SWE</td>
<td>0–1704</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0–2635</td>
<td></td>
</tr>
<tr>
<td>Minimum monthly humidity index**</td>
<td>P/PET</td>
<td>0.14–2.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.06–2.07</td>
<td></td>
</tr>
</tbody>
</table>

*The percentile from within the Tsuga heterophylla zone may exceed that for the entire region if T. heterophylla is clustered at one end of the gradient.
†The window width is a value used to create smoothed response surfaces. It is only given for variables used in the response surface analysis.
‡Calculated by interpolating monthly values.
§The water balance was modelled using a custom-written program, following Willmott et al. (1985). We assumed a uniform soil field capacity of 100 mm, and day length calculations followed Forsythe et al. (1995).
¶Sum of precipitation on months with mean temperature < −1 °C.
**The month with the lowest P/PET.
GDD5 (growing degree days on a 5 °C base) and MTCO (mean temperature of the coldest month). The other two sets of variables were the best three (AET, MTCO and AET/PET) and the best five (AET MTCO AET/PET GDD5 and P-PET: effective precipitation) variables identified by RTA. To make the calculations tractable, we limited the study area to the T. heterophylla zone plus 140 km immediately outside the zone, a distance chosen to encompass the area between the interior and coastal portions of the zone (n = 75,502 and 292,124 grid points within and outside of the T. heterophylla zone, respectively). Note that grid points correspond to the ecosystem-level map we used to infer presence and absence rather than actual presence/absence observations.

Goodness of fit of each model was assessed by comparing the predicted to the observed zone using the κ (kappa) statistic, a measure of the accuracy of a classification. We chose to use κ because the response variables of RTA and GM-SMAP are not continuous, precluding threshold-independent methods such as the receiver-operating characteristic (ROC plots; Fielding & Bell, 1997). κ is calculated as [Po−Pe]/[1−Pe], where Po is the proportion of grid points that agree between the predicted and observed zones and Pe is the proportion of grid points expected to agree by chance. κ ranges from −1 (the predicted and observed ranges are mirror images of each other) to 1 (the predicted zone exactly matches the observed zone). Estimating confidence intervals (CIs) of κ was difficult because of spatial autocorrelation and the fact that the large number of grid points unrealistically reduced the variance in κ. We estimated the 95% CI of κ to be c. ± 0.02 following Fleiss et al. (1969) after reducing the sample size to account for the fact that most spatial autocorrelation occurs at a scale of 10 km. Because of the problems with determining a CI, we compared κ-values cautiously.

The predictive capacities of the models were compared using a cross-validation procedure where the study area was divided into training data and testing data. Some studies have employed a leave-one-out cross-validation where one grid point is predicted based on a model trained on all other grid points (McCune et al., 2003). However, the fine spatial resolution and strong autocorrelation in the climatic data suggest that a leave-one-out cross validation would differ little from the model trained on the entire study area, and thus not adequately evaluate predictive capacity. We used a k-fold cross-validation strategy (Fielding & Bell, 1998) that more thoroughly evaluates predictive capacity by training models on small spatially stratified random subsamples. It involved training each of the three models on 100 sets of 1% subsamples (i.e. 3660 grid points) and computing κ on predictions made to the complement (99%) of grid points. This small, yet very representative, subsample was chosen as a rigorous test of the capacity of a model to predict in climates differing slightly from the training set. Grid points in each subsample were spaced ≥ 20 km to assure equal representation of the entire map and reduce autocorrelation among subsampled grid points. A model with good predictive capacity would have a consistently high cross-validation κ.

RESULTS

The regression tree shows that AET is the single most important variable at determining the T. heterophylla zone (Fig. 4). A single threshold of 443 mm correctly classifies 78.6% of the grid points, and adding MTCO and AET/PET increases the total percentage correctly classified to 84.0%. Two additional variables (GDD5 and P-PET) result in a marginal improvement (84.6%). No other variables meet the criteria of increasing the accuracy by 0.25%. The accuracy is much higher for low AET (outside the T. heterophylla zone) than for high AET (94.4% and 48.8%, respectively). Where AET ≤ 443 mm, adding five branches based on three additional variables (MTCO, AET/PET and P-PET) only increases accuracy slightly both within and outside the T. heterophylla zone. Where AET > 443 mm, adding a branch at an MTCO of −8.9 °C increases the accuracy to 66.4%, but further branching involving AET and GDD5 only marginally increases accuracy on this side of the tree.

Of the variable sets, the five-variable set fits the observed T. heterophylla zone significantly better than either three-variable set for the GM-SMAP and SRS models (Table 2). The increase in κ between these variable sets is 0.051 and 0.074 for GM-SMAP and SRS, respectively, which exceeds the estimated κ CI of ± 0.02 (see Methods). For RTA, κ increases only 0.024 between these variable sets, consistent with this model’s small increase in...
Bioclimatic modelling using Gaussian mixture distributions

Table 2 Results of resubstitution tests of three bioclimatic models predicting the *Tsuga heterophylla* zone

<table>
<thead>
<tr>
<th>Model</th>
<th>Variable set</th>
<th>Percent underprediction</th>
<th>Percent overprediction</th>
<th>(\kappa)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GM-SMAP</td>
<td>GDD5, MTCO, AET/PET</td>
<td>18.0</td>
<td>18.4</td>
<td>0.532</td>
</tr>
<tr>
<td></td>
<td>AET, MTCO, AET/PET</td>
<td>15.2</td>
<td>15.7</td>
<td>0.592</td>
</tr>
<tr>
<td></td>
<td>AET, MTCO, AET/PET, GDD5, P-PET</td>
<td>9.3</td>
<td>14.0</td>
<td>0.643</td>
</tr>
<tr>
<td>RTA</td>
<td>GDD5, MTCO, AET/PET</td>
<td>13.8</td>
<td>19.1</td>
<td>0.548</td>
</tr>
<tr>
<td></td>
<td>AET, MTCO, AET/PET</td>
<td>15.5</td>
<td>16.1</td>
<td>0.581</td>
</tr>
<tr>
<td></td>
<td>AET, MTCO, AET/PET, GDD5, P-PET</td>
<td>18.9</td>
<td>13.3</td>
<td>0.605</td>
</tr>
<tr>
<td>SRS</td>
<td>GDD5, MTCO, AET/PET</td>
<td>13.5</td>
<td>14.8</td>
<td>0.629</td>
</tr>
<tr>
<td></td>
<td>AET, MTCO, AET/PET</td>
<td>12.8</td>
<td>14.7</td>
<td>0.641</td>
</tr>
<tr>
<td></td>
<td>AET, MTCO, AET/PET, GDD5, P-PET</td>
<td>8.3</td>
<td>11.1</td>
<td>0.715</td>
</tr>
</tbody>
</table>

AET/PET = actual/potential annual evapotranspiration, GDD5 = growing degree-days on a 5 °C base, MTCO = mean temperature of the coldest month and P-PET = effective precipitation.

overall accuracy because of increased branching (Fig. 4). All models using the three best variables identified by RTA (AET, MTCO and AET/PET) have a better fit than the commonly used three-variable set (GDD5, MTCO and AET/PET), with \(\kappa\) differences of 0.012–0.060.

Of the three models using the five-variable set, SRS has the closest fit, and GM-SMAP has a better fit than RTA (Table 2). The improved fit of SRS over GM-SMAP (difference in \(\kappa\) = 0.072 for the five-variable set) is twice that for GM-SMAP over RTA (difference in \(\kappa\) = 0.038). The poorer fit of GM-SMAP is because of a 2.9% greater overprediction compared to SRS (Table 2). GM-SMAP and SRS overpredict in the same areas (blue areas on Fig. 5), but these areas are more contiguous in the GM-SMAP model and thus cover a larger area. In contrast, the poorer fit of RTA is because of a c. 9% greater underprediction compared to the other models. This underprediction occurs mainly in the northern valleys in the interior portion of the *T. heterophylla* zone (Fig. 5). All models overpredict in high elevation coastal areas of British Columbia and southeast Alaska with very steep terrain, and in the south-eastern portion of the interior range of *T. heterophylla*.

Cross-validation tests show that GM-SMAP has the greatest predictive capacity of the three models (Fig. 6). The median \(\kappa\) of 100 cross-validation tests is significantly greater for GM-SMAP (0.555) than for RTA (0.526) or SRS (0.480). The decrease in \(\kappa\) between models trained on the full study area and on 1% of the study area (cross validation) is much greater for SRS (0.235) than for RTA (0.079) or GM-SMAP (0.088).

![Figure 5](image)

Figure 5 Model predictions for the three bioclimatic models. The dashed line indicates the area within 140 km of the *T. heterophylla* range where bioclimatic models were trained and applied. All models were based on five climatic or water-balance variables (Table 2). The projection is Albers Equal Area, centred on British Columbia.
Examples of mapped cross-validation predictions show spatial patterns of prediction error (Fig. 7). All models greatly overpredict areas surrounding the interior portion of the *T. heterophylla* zone. SRS has the greatest overprediction, with much fine-scale variability, and it also underpredicts a large area in western Washington and Oregon. GM-SMAP and RTA show more contiguous patterns in overpredicted areas. The areas of overprediction are generally smaller for GM-SMAP than for RTA but the locations are similar.

**DISCUSSION**

Selection of climatic variables for species range prediction

RTA identified AET as the most predictive variable. Along with the climatic water deficit (which was highly correlated with AET/PET), AET has been shown to be a superior predictor of vegetation from local species distributions to biomes (Stephenson, 1998). AET integrates growing season temperature and available moisture (i.e. biologically available energy and water), and thus, species with different moisture and energy requirements should be patterned along the AET gradient. *T. heterophylla* is a species with one of the highest moisture requirements among conifers (Lassoie *et al*., 1986), hence its occurrence at high AET. In contrast, the variable set commonly used in bioclimatic models replaces AET with GDD5 (Table 2), which does not capture the interaction of energy and water availability (Stephenson, 1998). For *T. heterophylla*, this variable set performed poorly when compared with the best three or five variables identified by RTA, regardless of the model form (Table 2). These results demonstrate that RTA or other models that select among candidate variables (e.g. logistic regression) can be used to identify biologically meaningful variables for models that lack this capacity (Walker & Cocks, 1991). There is no automated method for selecting variables in SRS and GM-SMAP, and the alternative of comparing resubstitution and cross-validation results of all variable combinations would be an overwhelming task for large data sets.

Models using five variables were a large improvement over those using three variables, justifying the increased complexity of the two additional variables. This improvement was especially large for GM-SMAP and SRS whose fits increased two to three times more than that of RTA (Table 2). It is likely that GM-SMAP...
and SRS captured variable interactions better than RTA (Fig. 2). For example, the Gaussian mixture identified about twice the number of subclasses for the five-variable vs. three-variable sets (Table 3). In addition, for both GM-SMAP and SRS, five variables resulted in a smaller amount of underprediction than three variables (Table 2). Underprediction is likely a failure of the model algorithm, as it indicates that the model cannot predict areas where a species is known to occur. Overprediction, on the other hand, can have plausible ecological explanations (e.g. slow migration rates or interspecific competition) and thus is not as serious an error as underprediction (Svenning & Skov, 2004). The ecological interpretation of non-climatic factors affecting overprediction of *T. heterophylla* is the focus of another study (Gavin & Hu, unpublished data).

### Comparisons of bioclimatic models

Of the three models, GM-SMAP had the greatest predictive capacity (as determined by cross validation), which we attribute to the GM distribution for summarizing climatic relationships and the multiscale SMAP algorithm for making predictions. The GM distribution uses objective criteria (the MDL) to automatically determine the appropriate level of generalization (Grünwald, 2004). In contrast, RTA may be too general, not capturing the complexity of variable interactions because thresholds were based on a small number of perpendicular planes. On the other hand, SRS may be too specific, using a uniform climatic window that required enough data within the window to make a prediction. An advantage of the GM is that it includes variable interactions (using a variance–covariance matrix for each subclass) and adapts to sharp or diffuse transitions of species abundance on climatic gradients. The latter property means that GMs can identify abrupt climatic thresholds but also interpolate across poorly sampled regions in climate space in the same model (Fig. 2a). For example, the GM identified 61 subclasses when trained on the *T. heterophylla* zone over the entire study area (Table 3), and an average of five subclasses when trained on cross-validation subsamples. The smaller number of subclasses for the smaller cross-validation subsamples indicated greater generalization and interpolation within climate space when using sparser data. We suggest that this adaptability of the GM method increased the predictive capacity of GM-SMAP relative to RTA and SRS.

The use of the MDL in the GM fitting algorithm is highly suitable for predictive modelling of species ranges as it automates model selection, resulting in the 'best' model as a trade-off between fit and generalization (Rushton et al., 2004). The MDL or other information-theoretic approaches [e.g. the Akaike Information Criteria (AIC) Burnham & Anderson, 2002] can be used not only to parameterize models of a particular form (e.g. GM distributions), but also to compare models with diverse forms. However, because SRS, as applied in this study, is a smoother of the data instead of a parameterized model (see below), it could not be described by such criteria. Therefore, we relied on cross validation as a practical demonstration of a model's robustness and ability to predict in areas outside the training data set.

GM-SMAP makes predictions from the GM 'signature' using a multiscale algorithm (SMAP) that classifies each grid point in the context of the climate of neighbouring points, minimizing prediction of small outlier areas. SMAP increased the $\kappa$ of cross-validation tests by an average of 0.015 when compared to a point-by-point maximum likelihood method (not shown) that used the same GM distribution. Thus, the SMAP algorithm was partly responsible for the improved fit of GM-SMAP over RTA and SRS in cross-validation (difference in $\kappa$ c. 0.03 and 0.075, respectively; Fig. 6). In contrast, the spatially unrealistic SRS cross-validation result (Fig. 7) may be partly due to predictions in each grid point made independently of the neighbouring grid points. Similarly, improved fit because of the SMAP algorithm has been found in remote sensing studies where it outperformed maximum-likelihood classification (McCaughey & Engel, 1995) and artificial neural networks (Michelson et al., 2000).

In contrast to RTA and GM-SMAP, SRS has no automated method to guard against overfitting, and the method is difficult to implement. SRS is better described as a 'smoother' than as a statistical model because it does not generalize the training data into a compressed format. The entire training data set is required to make each prediction, requiring long computer run times. This lack of generalization makes it difficult to use a selection method such as MDL to aid model selection (e.g. window width of the smoothing function). Instead, the window width used in SRS is set a priori, and determining the optimal window for each variable via an iterative trial-and-error procedure would be prohibitive for large data sets. In this study, the window size (20% of the central 95% range of each variable) was likely too specific. These issues highlight a significant advantage of the GM-SMAP model: no subjective decisions are required once the predictor variables are chosen.
The characteristics of the GM-SMAP model and its capacity to predict *T. heterophylla* suggest that it performs well in other settings. However, several other recently introduced models may perform equally well. For example, multivariate adaptive regression splines (MARS) and general additive models (GAM) are similar to or more competitive than RTA (Prasad & Iverson, 2000; Moisen & Frescino, 2002; Banks et al., 2003; Munoz & Felicisimo, 2004). MARS appears to be superior to many models at fitting data where only a portion of the variables is explanatory (Banks et al., 2003), and artificial neural networks are very predictive, though dependent on the characteristics of the species distribution (Segurado & Araujo, 2004). Further evaluation of GM-SMAP should include comparisons with these models and with a variety of training data sets.

ACKNOWLEDGEMENTS

This research was funded by the US National Science Foundation (DEB 02–12917) and the Packard Foundation. The authors thank two anonymous reviewers for comments on the manuscript.

REFERENCES


**BIOSKETCHES**

**Daniel Gavin** is a postdoctoral associate in the Department of Plant Biology at the University of Illinois, Champaign-Urbana, IL. He is interested in Quaternary palaeoecology, biogeography, fire history and forest ecology.

**Feng Sheng Hu** is an associate professor in the Departments of Plant Biology and Geology at the University of Illinois. He is interested in climatic change and ecosystem response.