Potential habitat distribution for the freshwater diatom *Didymosphenia geminata* in the continental US

Sunil Kumar[1*], Sarah A Spaulding[2,3], Thomas J Stohlgren[3], Karl A Hermann[2], Travis S Schmidt[4], and Loren L Bahls[5]

The diatom *Didymosphenia geminata* is a single-celled alga found in lakes, streams, and rivers. Nuisance blooms of *D geminata* affect the diversity, abundance, and productivity of other aquatic organisms. Because *D geminata* can be transported by humans on waders and other gear, accurate spatial prediction of habitat suitability is urgently needed for early detection and rapid response, as well as for evaluation of monitoring and control programs. We compared four modeling methods to predict *D geminata*'s habitat distribution; two methods use presence–absence data (logistic regression and classification and regression tree [CART]), and two involve presence data (maximum entropy model [Maxent] and genetic algorithm for rule-set production [GARP]). Using these methods, we evaluated spatially explicit, bioclimatic and environmental variables as predictors of diatom distribution. The Maxent model provided the most accurate predictions, followed by logistic regression, CART, and GARP. The most suitable habitats were predicted to occur in the western US, in relatively cool sites, and at high elevations with a high base-flow index. The results provide insights into the factors that affect the distribution of *D geminata* and a spatial basis for the prediction of nuisance blooms.

Front Ecol Environ 2009; 7, doi: 10.1890/080054

Environmental change in North America has reinforced the importance of habitat modeling, to determine the habitat preferences and potential geographic distributions of invasive species in terrestrial (Stohlgren *et al.* 2006) and aquatic (Williamson *et al.* 2008) systems. The diatom *Didymosphenia geminata* is a single-celled alga (Bacillariophyceae; Figure 1) that is becoming increasingly prevalent in North America (Spaulding and Elwell 2007) and is invasive in New Zealand (Kilroy *et al.* 2008). This diatom has been reported in the western US for over 100 years, but more extensive, nuisance growths have recently become common; nuisance growths are also appearing with greater frequency in the eastern US. In New Zealand, this species was initially discovered on the South Island in 2004, and it is now present in over 21 rivers (Duncan 2007), and forms large growths at several sites. *D geminata* has the potential to generate serious ecological and economic impacts in both these countries. Unpublished studies in the US and published studies in New Zealand (eg Kilroy *et al.* 2006) indicate large increases in algal biomass at sites impacted by *D geminata*, and shifts in algal species composition. There are also differences in the major invertebrate groups between non-impacted and impacted sites (Kilroy *et al.* 2006). Predictive modeling provides an opportunity to examine the role of specific environmental variables that may be associated with the distribution of a species at various spatial scales and can help to determine appropriate management actions. Our objectives were to: (1) predict the potential suitable habitats for *D geminata*; (2) compare four species distribution modeling techniques for predicting suitable habitats for *D geminata*; and (3) determine *D geminata*'s response to bioclimatic, topographic, geologic, and hydrologic variables in the continental US.

[1]*Natural Resource Ecology Laboratory, Colorado State University, Fort Collins, CO* [*]*(sunil@nrel.colostate.edu);* [2]*Environmental Protection Agency, Region 8, Denver, CO;* [3]*US Geological Survey, Fort Collins Science Center, Fort Collins, CO;* [4]*US Geological Survey, Mineral Resources Program, Denver, CO;* [5]*Hannaea, Helena, MT*

Methods

Presence–absence data

We compiled data from several sources to develop a potential distribution map of *D geminata* in the US (WebTable 1). Absolute and relative abundance estimates of diatom cells were converted to presence–absence, based on survey data from over 4750 samples. "Presence" was defined as the positive recording of a cell; abundance was not considered. Likewise, we defined "absence" as the lack of observed *D geminata* cells during a 300-diatom cell count using an oil immersion 100x objective with light microscopy. After removing multiple records, so that only one record per 1 km x 1 km cell remained, we found that 308 presence points and 2724 absence points were left (Figure 2a).

Environmental variables

We initially considered 39 spatially explicit environmental variables, representing climate (eg temperature, precipitation, radiation, growing degree days, number of frost

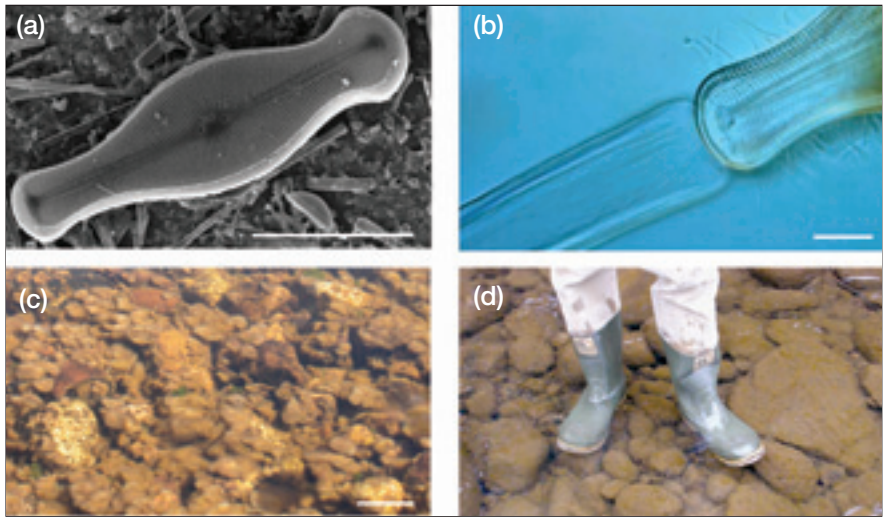

Figure 1. *(a) Scanning electron micrograph of* D geminata. *Scale bar equals 0.05 mm. (b) Light micrograph of portion of living cell, showing extracellular stalk. Scale bar equals 0.01 mm. (c) Image of actively growing colonies attached to cobbles in a stream. Scale bar equals approximately 10 cm. (d) Felt-soled waders in a shallow stream with 100% coverage of the substrate by* D geminata *and stalks. The diatom cells are capable of surviving transport on anglers' equipment.*

days, and humidity), topography (elevation, slope, and aspect), land-use and land-cover types, enhanced vegetation index, bedrock geology, and hydrology (eg base-flow index, flow accumulation, and flow direction) for the continental US (WebTable 2). We calculated 19 bioclimatic variables (www.worldclim.org/bioclim.htm; Nix 1986) – which are biologically more meaningful than just annual means for defining the ecophysiological tolerances of a species – using ARC AML script (MkBCvars.AML; www.worldclim.org/mkBCvars.aml; Hijmans 2006) using the Daymet climate dataset (www.daymet.org/; 1-km spatial resolution; 1980–1997; WebTable 2). Variations in vegetation conditions were represented by the moderate resolution imaging spectroradiometer (MODIS) enhanced vegetation index (EVI; WebTable 2), an optimized vegetation index that captures changes in biomass. All geographic information system (GIS) layers representing environmental variables were resampled to a resolution of 1 km, to match the 19 bioclimatic variables. We conducted all GIS analyses using Environmental Systems Research Institute's (ESRI, Redlands, CA) ARC GIS, version 9.1. Multicollinearity was tested by examining cross-correlations among all the variables. Only one variable from a set of highly correlated variables was included in the analyses (ie Pearson correlation coefficient $\geq \pm 0.80$), based on its potential ecological relevance to *D geminata*'s distribution and for ease of interpretation. For example, maximum annual temperature, minimum annual temperature, number of frost days, growing degree days, and humidity were highly correlated; we included growing degree days and dropped others. Thus, the final number of variables considered for all four modeling methods was reduced to 26 (WebTable 2).

Modeling methods

We compared four different modeling methods for predicting potential habitat distribution for *D geminata*, including two presence–absence and two presence-only methods. We implemented the two presence–absence methods – stepwise multiple logistic regression and classification and regression trees (CART) – using SYSTAT statistical software (version 12; Systat Software Inc 2007, San Jose, CA). The best logistic regression model was selected, based on lowest Akaike's information criterion (AIC) values, and alpha = 0.05 was used to determine the significance of the predictors. Presence-only methods included a fairly recently introduced method called Maxent (maximum entropy modeling; Phillips *et al.* 2006), and the widely used GARP (genetic algorithm for rule-set prediction; Stockwell and Noble 1992).

Maxent is a machine learning method (version 3.1; www.cs.princeton.edu/~schapire/maxent/) and is based on the maximum entropy principle. It assesses the probability distribution of a species by estimating the probability distribution of maximum entropy (Phillips *et al.* 2006). A recent model comparison by Elith *et al.* (2006) ranked Maxent as the best-performing model algorithm out of a total of 16 different modeling methods; however, these comparisons were limited to terrestrial plants, birds, bats, and reptiles (Elith *et al.* 2006). There are no comparative studies on how these methods would perform in predicting the spatial distribution of an aquatic species. Most of the studies on aquatic species distribution have used only one modeling approach (eg Drake and Bossenbroek 2004). We ran Maxent using the linear, quadratic, product threshold, and binary features (for details, see Phillips *et al.* 2006). The jackknife variable importance feature in Maxent was used to assess the relative importance of the environmental predictors in the model.

GARP models were developed using a desktop version of GARP (http://nhm.ku.edu/desktopgarp/index.html). GARP uses a set of rules to relate species presence data to the prevailing environmental conditions. Since GARP predictions are stochastic, we implemented the best-subset model selection procedure (Peterson and Shaw 2003). We generated 200 binary models (1's for predicted pixels and 0's for unpredicted pixels) using a 0.01 convergence limit, 1000 maximum iterations, and allowing the use of atomic, range, negated range, and logit rules. A best subset of 10 models, based on 5% intrinsic omission of training localities threshold, was selected. Final GARP prediction was obtained by combining the ten best subset models in ARC Map version 9.1, in which the value of pixels varied from 0–10, with "0" representing the pixels

that were not predicted by any of the models (ie absence of *D geminata*) and "10" representing the pixels that were predicted as showing the presence of *D geminata* by all ten models.

Model development and validation

We randomly selected 308 absence samples from the 2724 available, to match the number of spatially unique presence records (308) and maintain an intermediate level of sampling prevalence (proportion of samples representing species presence) for logistic regression and CART models (Fielding and Bell 1997; McPherson *et al.* 2004). We randomly partitioned these 308 presence and 308 absence samples into training (50%) and validation (50%) datasets ("split sample" approach), thus creating a quasi-independent dataset for model validations (Guisan and Hofer 2003). The training dataset (n = 308; 154 each for presence and absence; Figure 2a) was used to develop models, using all four modeling methods (only presence records, 154, were used in Maxent and GARP), and the remaining data were used for validation. All presence samples (308) were used in the final model, which was obtained by the highest ranked model, the Maxent model (Table 1; Figure 2b).

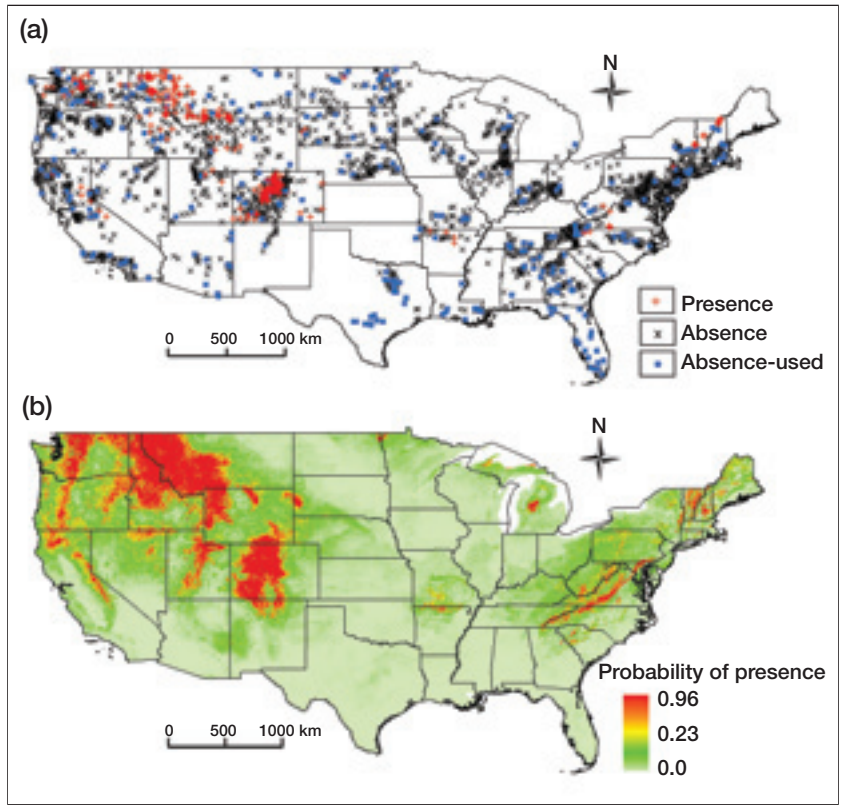

Figure 2. *(a) Spatial distribution of* D geminata*'s presence (308) and absence (2724) locations; absence-used are randomly selected absence records (308) used in logistic regression and CART models. (b) Predicted probability for* D geminata*'s presence, based on the best model via all the presence records (ie Maxent).*

Modeling methods were compared, on the basis of their performances; they were evaluated using four threshold-dependent measures, including sensitivity, specificity, correct classification rate (CCR) or overall accuracy, and Cohen's maximized Kappa (K) – as well as by a threshold-independent measure – area under the receiver operating characteristic (ROC) curve (AUC; for details, see Fielding and Bell 1997). Sensitivity, also called true positive rate, is the fraction of all presences correctly classified as "presence". Specificity is the fraction of all absences correctly classified as "absence". CCR was calculated by dividing the sum of correctly classified presence and absence records by the total number of samples (Fielding and Bell 1997). The maximized Kappa statistic is obtained by plotting sensitivity and specificity against different thresholds to define decision thresholds where the two curves cross. Kappa ranges from –1 to 1, where 1 represents a perfect agreement, whereas values less than 0 indicate model performance no better than that produced by chance (also described as $K < 0.40$, poor; $0.40 < K < 0.75$, good; and $K > 0.75$, excellent performance; Fielding and Bell 1997). AUC quantifies the model performance at all possible thresholds and is obtained by plotting sensitivity (y axis) against 1 – specificity (called false positive; x axis). AUC varies from 0.5 for models performing no better than that produced by chance, or with no dis-

Table 1. Model validation and evaluation summary for *D geminata*

Modeling method	*Training data* (154 presence/154 absence)*					*Validation data (154 presence/154 absence)*				
	AUC	Sen	Spe	K	CCR	AUC	Sen	Spe	K	CCR
Maxent	0.95	0.95	0.85	0.80	0.90	0.92	0.83	0.91	0.74	0.87
Logistic regression	0.90	0.73	0.91	0.64	0.82	0.91	0.88	0.81	0.70	0.85
CART	0.93	0.88	0.90	0.77	0.89	0.86	0.83	0.76	0.59	0.80
GARP	0.83	0.94	0.64	0.58	0.79	0.82	0.88	0.70	0.57	0.79

Notes: AUC = area under receiver operating characteristic (ROC) curve. Sen = sensitivity, the fraction of all presences correctly classified as "presence". Spe = specificity, the fraction of all absences correctly classified as "absence". K = maximized Cohen's Kappa. CCR = correct classification rate or overall accuracy. *Only presence data were used for training Maxent and GARP models; however, both presence and absence data were used to calculate five model evaluation statistics.

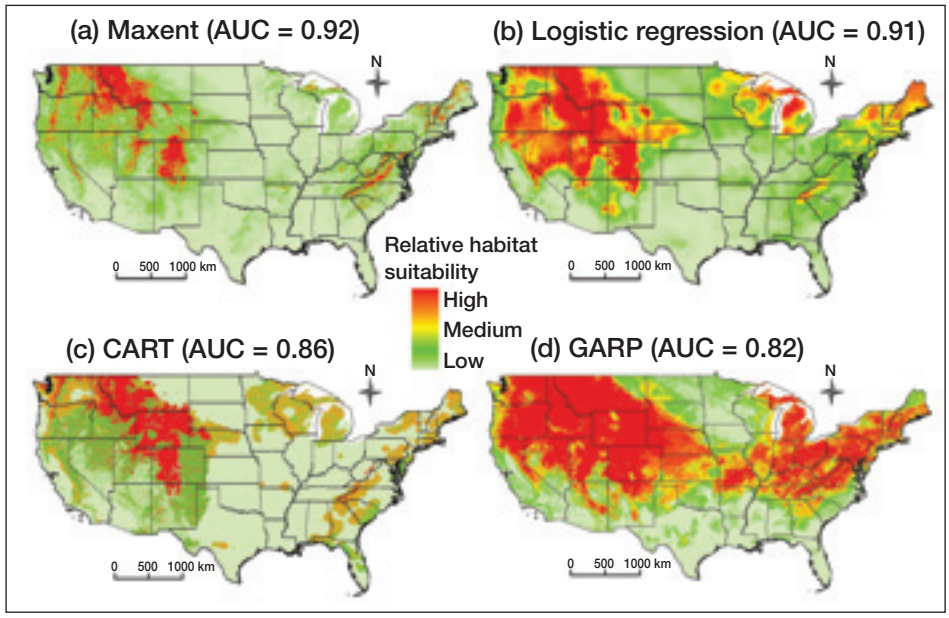

Figure 3. *Predicted habitat suitability for* D geminata *based on (a) Maxent; (b) logistic regression; (c) classification and regression trees (CART); and (d) genetic algorithm for rule-set prediction (GARP) modeling algorithms. AUC is area under receiver operating characteristic (ROC) curve.*

crimination ability, to 1.0 for models performing with perfect discrimination.

Results

Model comparisons based on five different evaluation statistics showed that the Maxent model was the best performer, with a validation AUC of 0.92, Kappa of 0.74, and overall accuracy of 0.87, followed by logistic regression (Table 1; Figure 3a). The other two modeling methods, CART and GARP (in order of their ranks based on AUC and K; Table 1), performed poorly when compared with Maxent and logistic regression models (Figure 3). The Maxent model results revealed the most suitable areas for *D geminata* in the western US; these were primarily in western Montana, northern Idaho, northwestern Wyoming, and the Colorado Rocky Mountains (Figure 2b). Spatially, models other than Maxent predicted more areas with highly suitable habitats (Figure 3).

Base-flow index was one of the best predictors of *D geminata*'s presence and was selected in all four models (WebTable 2). The jackknife test of variables' importance in the Maxent model indicated that mean temperature during the warmest quarter and base-flow index were two of the best predictors of potentially suitable habitat for *D geminata*, with 30.3% and 14.5% contributions, respectively (WebTable 2). Environmental conditions varied widely at locations where *D geminata* was present (WebTable 2). It was found at elevations ranging from 65 m to 3853 m, and in regions where the average annual temperature varied from –5°C to 16°C, and where average annual precipitation ranged from 198 mm to 3253 mm (WebTable 2). Our results suggest that there is a relatively higher probability of finding *D geminata* in headwater streams with a higher base-flow index and at higher elevations, and therefore with cooler climates (Figure 4). This is consistent with this species' habitat preferences, which have historically been reported as cold, fast-flowing, low nutrient streams (Spaulding and Elwell 2007). This analysis shows that a large component of the distribution of *D geminata* can be attributed to climatic factors alone, at least at a continental scale.

The best logistic regression model explained 60% of the variation in *D geminata*'s presence or absence (Naglekerke's R^2 = 0.604) and included four environmental predictors: annual mean temperature (–ve), isothermality (+ve), precipitation seasonality (–ve), and base-flow index (+ve), all significant at alpha = 0.05. The CART model explained 64% of the variation in *D geminata*'s presence–absence (proportional reduction in error [PRE] = 0.643), with eight terminal nodes and seven predictor variables. Elevation and base-flow index were the two most important predictor variables in the CART model, followed by precipitation in the driest month, mean MODIS EVI, mean temperature of the driest quarter, flow accumulation, and compound topographic index (WebTable 2).

Discussion and conclusions

Historically, the species composition of diatoms in freshwaters was thought to be strongly influenced by water chemistry variables, including concentration of phosphorus, nitrate, trace metals, and dissolved organic carbon, as well as pH, specific conductance, and other variables (Stoermer and Smol 1999; Potapova and Charles 2007). The relationship between diatom species and dissolved solutes has been the basis for the usefulness of diatoms in aquatic assessment and paleolimnology, including paleoreconstruction of climate (eg Smol and Douglas 2007). Although temperature is considered to be a less influential variable than some others (Anderson 2000), diatom species composition has been linked to climatic fluctuations, including surface-water temperature (Verleyen *et al.* 2003; Potapova and Winter 2006; Vyverman *et al.* 2007). In studies where both air and surface-water temperatures were evaluated, the relationship between diatom species composition and air temperature has been more robust than that between species composition and surface-water temperature (Joynt and Wolfe 2001; Bloom

et al. 2003). Here, we used air temperature, along with several other GIS-derived variables (Nix 1986), in a first attempt to model diatom habitat distribution at a continental scale. It is clear that our findings need to be tested further, by application to new systems in other parts of the world, before they can be broadly generalized.

We evaluated four models and found that, for *D geminata*, the Maxent model performed noticeably better than the others (Table 1). The better performance of the Maxent model can be attributed to the complexity of its underlying algorithm, as compared with other modeling methods, and its ability to model the complex shapes of species' responses to environmental factors. We conclude that it is advisable to compare modeling approaches, particularly because techniques that are successful for one species may not be successful for others, as has been shown for habitat specialists and generalists by Evangelista *et al.* (2008). Our results also suggest that in the case of aquatic organisms, presence-only models such as Maxent can perform as well as presence–absence models (eg logistic regression or CART; Elith *et al.* 2006). This finding could be more important for modeling distributions of many introduced species, for which data are often limited to presence-only. Finally, we found that models varied in terms of spatial predictions; these differences could be due to (1) the inclusion of presence-only data (Maxent, GARP) versus the presence–absence data (logistic regression, CART), (2) differences in GIS variables included in different models (WebTable 2), and (3) the underlying assumptions and complexity of the algorithms of different models (Elith *et al.* 2006). For example, logistic regression models considered only linear responses, whereas the Maxent model algorithm considered linear, non-linear, and interaction effects (Phillips *et al.* 2006).

The model was able to successfully predict *D geminata*'s potential habitat distribution in the continental US, without the use of water chemistry data. Of course, water chemistry variables and climate are related, but we accounted for a high degree of variance in distribution based on air temperature alone. The importance of the base-flow index in our results is supported by the observation that *D geminata* is common in regulated rivers (Kirkwood *et al.* 2008) and lake-fed rivers with stable flow regimes (Kilroy *et al.* 2008). Although we have examined bioclimatic factors that explain the presence–absence of *D geminata* at a continental scale, we recognize that other factors may be relevant to the range expansion and formation of nuisance blooms by this species. Kilroy *et al.* (2008) established that *D geminata* is able to survive in damp conditions for more than 60 days, and viable cells were documented within felt-soled waders worn by anglers (Figure 1d). There is strong evidence that *D geminata* is spread by humans and their activities, particularly in the case of its introduction to New Zealand.

The factors influencing regional-scale distribution may be more appropriately addressed by incorporating information on anthropogenic factors that can affect *D geminata*'s distribution and by including spatially explicit data on water chemistry. Anthropogenic factors, such as recreational use, can influence the spread of this diatom (eg Bossenbroek *et al.* 2001). Modeling efforts in New Zealand have shown that temperature, stability (hydrological and substrate), solar radiation, and pH were the most important factors in determining where *D geminata* will form nuisance blooms (Kilroy *et al.* 2008). Water chemistry variables could be direct predictors of *D geminata*'s abundance; however, these data are not yet available in GIS format at regional or continental scales. The next step is to examine the distribution of nuisance blooms and their relationship to anthropogenic factors and water chemistry.

We have established the climatic range for this species in the continental US. The discovery that mean temperature during the warmest quarter was the most important factor in influencing distribution implies that the distribution of this species will be very sensitive to climatic change, particularly in the western US. The importance of base-flow index suggests that drought and water release from reservoirs could play a role in the development of nuisance blooms, and that the potential control of water flow could serve as a basis for management actions. Furthermore, the response of this

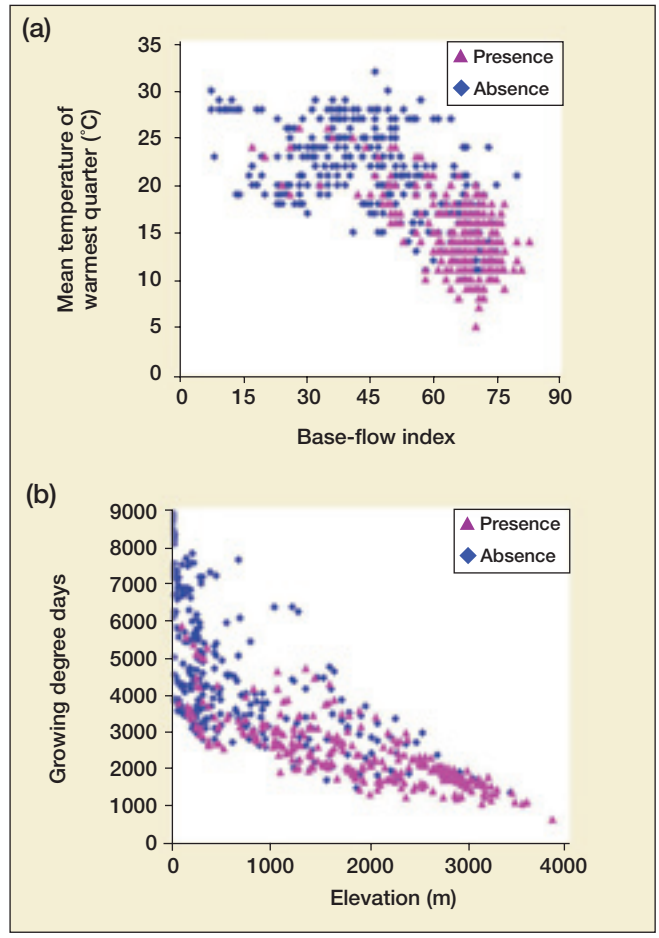

Figure 4. *Scatter plots of (a) base-flow index versus mean temperature during the warmest quarter (°C), and (b) elevation versus growing degree days, with known presence and absence locations of* D geminata *in the continental US.*

species to climate change and watershed alteration is an example of the ability of stream organisms to adapt to the effects of environmental change (Williamson *et al.* 2008). We hope that our findings will be useful in controlling the spread of *D geminata* and managing the size of its blooms, as well as for minimizing its impacts on fisheries, water supplies, tourism, biodiversity, and aesthetic values.

Acknowledgements

We thank M Bothwell, C Kilroy, C Vieglais, and M Potapova for helpful discussions. This work is based on the development of the EMAP program in Corvallis, OR, and the work of A Herlihy, P Kaufmann, P Larson, S Paulson, D Peck, and J Stoddard. N Gillett and K Manoylov provided additional information. We thank EPA Region 8, the Natural Resource Ecology Laboratory at Colorado State University, and USGS Fort Collins Science Center for funding and logistical support. TJS and SK acknowledge funding for data analysis from NASA grant NRA-03-OES-03. Any use of trade, product, or firm names is for descriptive purposes only and does not imply endorsement by the US Government.

References

Anderson NJ. 2000. Diatoms, temperature and climatic change. *Eur J Phycol* **35**: 307–14.

Bloom AM, Moser KA, Porinchu DF, *et al.* 2003. Diatom-inference models for surface-water temperature and salinity developed from a 57-lake calibration set from the Sierra Nevada, California, USA. *J Paleolimnol* **29**: 235–55.

Bossenbroek JM, Kraft CE, and Nekola JC. 2001. Prediction of long-distance dispersal using gravity models: zebra mussel invasion of inland lakes. *Ecol Appl* **11**: 1778–88.

Drake JM and Bossenbroek JM. 2004. The potential distribution of zebra mussels in the United States. *BioScience* **54**: 931–41.

Duncan M. 2007. New Zealand-wide surveys in November 2006, February 2007 and May 2007 for the presence of the non-indigenous freshwater diatom *Didymosphenia geminata* in high risk sites. Christchurch, New Zealand: National Institute of Water and Atmospheric Research Ltd. Client Report CHC2007-053.

Elith J, Graham CH, Anderson RP, *et al.* 2006. Novel methods improve prediction of species' distributions from occurrence data. *Ecography* **29**: 129–51.

Evangelista P, Kumar S, Stohlgren TJ, *et al.* 2008. Model selection for predicting a habitat gerneralist (*Bromus tectorum*) and a specialist (*Tamarix chinensis*) invasive plant species in Grand Staircase Escalante National Monument, Utah, USA. *Divers Distrib* **14**: 808–17.

Fielding AH and Bell JF. 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environ Conserv* **24**: 38–49.

Hijmans RJ. 2006. MkBCvarsAML version 2.3. www.worldclim.org/mkBCvars.aml. Viewed 29 Oct 2008.

Guisan A and Hofer U. 2003. Predicting reptile distributions at the mesoscale: relation to climate and topography. *J Biogeogr* **30**: 1233–43.

Joynt III EH and Wolfe AP. 2001. Paleoenvironmental inference models from sediment diatom assemblages in Baffin Island lakes (Nunavut, Canada) and reconstruction of summer water temperature. *Can J Fish Aquat Sci* **58**: 1222–43.

Kilroy C, Biggs B, Blair N, *et al.* 2006. Ecological studies on *Didymosphenia geminata*. Christchurch, New Zealand: National Institute of Water and Atmospheric Research Ltd. Report CHC2005-123, Project MAF05505.

Kilroy C, Snelder TN, Floerl O, *et al.* 2008. A rapid technique for assessing the suitability of areas for invasive species applied to New Zealand's rivers. *Divers Distrib* **14**: 262–72.

Kirkwood A, Jackson LJ, and McCauley E. 2008. *Didymosphenia geminata* and bloom formation along the south-eastern slopes of the Canadian Rockies. In: Bothwell ML and Spaulding SA (Eds). Proceedings of the 2007 International Workshop on *Didymosphenia geminata*. Nanaimo, Canada: Fisheries and Oceans Canada. Canadian Technical Report of Fisheries and Aquatic Sciences 2795.

McPherson JM, Jetz W, and Rogers DJ. 2004. The effects of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact? *J Appl Ecol* **41**: 811–23.

Nix HA. 1986. A biogeographic analysis of Australian elapid snakes. In: Longmore R (Ed). Australian flora and fauna series 8. Canberra, Australia: Australian Government Publishing Service.

Peterson AT and Shaw JJ. 2003. *Lutzomyia* vectors for cutaneous leishmaniasis in southern Brazil: ecological niche models, predicted geographic distributions, and climate change effects. *Int J Parasitol* **33**: 919–31.

Phillips SJ, Anderson RP, and Schapire RE. 2006. Maximum entropy modeling of species geographic distributions. *Ecol Model* **190**: 231–59.

Potapova M and Charles D. 2007. Diatom metrics for monitoring eutrophication in rivers of the United States. *Ecol Indic* **7**: 48–70.

Potapova MG and Winter DM. 2006. Use of nonparametric multiplicative regression for modeling diatom habitat: a case study of three *Geissleria* species from North America. In: Ognjanova-Rumenova N and Manoylov K (Eds). Advances in phycological studies. Sofia, Bulgaria: Pensoft.

Smol JP and Douglas MSV. 2007. From controversy to consensus: making the case for recent climatic change in the Arctic using lake sediments. *Front Ecol Environ* **5**: 466–74.

Spaulding SA and Elwell L. 2007. Increase in nuisance blooms and geographic expansion of the freshwater diatom *Didymosphenia geminata:* recommendations for response. Denver, CO: US Environmental Protection Agency Region 8. Open File Report 2007-1425. www.epa.gov/region8/water/didymosphenia/. Viewed 29 Oct 2008.

Stockwell DRB and Noble IR. 1992. Induction of sets of rules from animal distribution data: a robust and informative method of data analysis. *Math Comput Simulat* **33**: 385–90.

Stoermer EF and Smol JP. 1999. The diatoms: applications for the environmental and Earth sciences. Cambridge, UK: Cambridge University Press.

Stohlgren TJ, Barnett D, Flather C, *et al.* 2006. Species richness and patterns of invasion in plants, birds, and fishes in the United States. *Biol Invasions* **8**: 427–47.

Verleyen E, Hodgson DA, Vyverman W, *et al.* 2003. Modelling diatom responses to climate induced fluctuations in the moisture balance in continental Antarctic lakes. *J Paleolimnol* **30**: 195–215.

Vyverman W, Verleyen E, Sabbe K, *et al.* 2007. Historical processes constrain patterns in global diatom diversity. *Ecology* **88**: 1924–31.

Williamson CE, Dodds W, Kratz TK, and Palmer MA. 2008. Lakes and streams as sentinels of environmental change in terrestrial and atmospheric processes. *Front Ecol Environ* **6**: 247–54.

WebTable 1. Documentation of samples included in the models. To be included in this dataset, the identification of *D geminata* was required to be made within US Geological Survey (USGS) or Environmental Protection Agency (EPA) programs, or by Loren L Bahls (LLB), Travis S Schmidt (TSS), or Sarah A Spaulding (SAS).

Data source	*Years*	*Total number of samples*	*Number of samples with* D geminata *present*	*Citation*	*Sample archive location*
USGS National Water Quality Assessment (NAWQA)	1993–2007	3450	100	Potapova and Charles (2007)	Academy of Natural Sciences of Philadelphia, Philadelphia, PA
EPA Western Environmental Monitoring and Assessment Program (EMAP)	2000–2004	850	56	Stoddard *et al.* (2005)	California Academy of Sciences, San Francisco, CA
EPA Regional Environmental Monitoring and Assessment Program (REMAP)	1994–1995	108	17	Pollard and Yuan (2006)	California Academy of Sciences, San Francisco, CA
USGS Central Colorado Assessment Project (CCAP)	2005	59	59	Unpublished data	USGS, Denver, CO
Hannaea, Montana Diatom Database	1977–2005	127	127	Bahls (2004)	University of Montana Herbarium, Missoula, MT
Other – samples submitted to SAS by state and federal agencies, non-profit organizations, and the public	1976–2006	66	66	Spaulding and Elwell (2007)	Institute of Arctic and Alpine Research (INSTAAR) Diatom Collection, University of Colorado, Boulder, CO

References

Bahls L. 2004. Northwest diatoms: a photographic catalogue of species in the Montana Diatom Collection, vol 1. Helena, MT: Hannaea.

Pollard AI and Yuan L. 2006. Community response patterns: evaluating benthic invertebrate composition in metal-polluted streams. *Ecol App* **16**: 645–655.

Potapova M and Charles D. 2007. Diatom metrics for monitoring eutrophication in rivers of the United States. *Ecol Indic* **7**: 48–70.

Spaulding SA and Elwell L. 2007. Increase in nuisance blooms and geographic expansion of the freshwater diatom *Didymosphenia geminata:* recommendations for response. Denver, CO: US Environmental Protection Agency Region 8. Open File Report 2007-1425. www.epa.gov/region8/water/didymosphenia/. Viewed 29 Oct 2008.

Stoddard JL, Peck DV, Olsen AR, *et al.* 2005. Ecological assessment of western streams and rivers. Corvallis, OR: US EPA.

WebTable 2. Bioclimatic profile of *D geminata* based on all 308 presence locations in the continental US and environmental variables included in different modeling methods.

Environmental variable	*Percent contribution in Maxent model*	*Mean*	*SD*	*Minimum*	*Maximum*
[1]Mean temperature of warmest quarter (BIO10; °C) §	30.3	14.68	3.68	5.00	26.00
[2]Base-flow index (%) §, γ, ¥	14.5	65.94	10.28	17.00	83.00
[1]Frequency of precipitation (number of wet days/total days) §	9.2	0.28	0.07	0.11	0.47
[3]Geology §	8.0	na	na	na	na
[4]Elevation (m) §, ¥	7.6	1837.92	926.54	65.00	3853.00
[4]Flow accumulation (area in m^2) §, ¥	6.0	448.27	2466.08	0.00	25656.00
[1]Growing degree days (degree-days) §	5.1	2420.52	913.82	603.25	5823.85
[1]Annual precipitation event size (cm/day) §	4.1	0.77	0.26	0.39	2.19
[1]Isothermality (BIO3) §, γ	3.0	38.01	3.39	25.00	48.00
[5]Range in MODIS enhanced vegetation index (EVI) §	2.9	2584.64	820.23	376.82	5392.83
[1]Temperature seasonality (SD x 100) (BIO4) §	2.7	784.32	88.81	446.00	1257.00
[1]Radiation (MJ per m^2 per day) §	1.3	14.68	1.56	10.83	17.86
[1]Precipitation seasonality (CV) (BIO15) §, γ	0.8	295.00	144.38	80.00	750.00
[6]Land-use and land-cover types §	0.7	na	na	na	na
[4]Northness (cos[aspect]) §	0.7	–0.01	0.70	–1.00	1.00
[1]Precipitation of driest quarter (BIO17; cm) §	0.7	121.70	57.07	18.33	328.89
[1]Precipitation of wettest quarter (BIO16; cm) §	0.6	301.06	175.37	71.94	1389.83
[4]Flow direction §	0.5	28.01	36.18	1.00	128.00
[4]Eastness (sin[aspect]) §	0.3	0.01	0.71	–1.00	1.00
[4]Compound topographic index §, ¥	0.3	529.54	316.03	74.00	1582.00
[1]Mean temperature of wettest quarter (BIO8; °C) §	0.2	3.48	8.27	–11.00	23.00
[5]Mean of MODIS EVI §, ¥	0.2	836.38	597.78	95.10	3270.14
[1]Mean temperature of driest quarter (BIO9; °C) §, ¥	0.2	7.06	9.06	–13.00	26.00
[1]Mean diurnal range in temperature (BIO2; °C) §	0.1	13.96	1.89	9.00	18.00

Continued

WebTable 2. – *continued*

Environmental variable	*Percent contribution in Maxent model*	*Mean*	*SD*	*Minimum*	*Maximum*
[1]Precipitation of warmest quarter (BIO18; cm) §	0.1	159.03	69.36	30.00	420.00
[4]Slope (degrees) §	0.0	4.86	4.13	0.00	25.56
[1]Precipitation of driest month (BIO14; cm) ¥	–	33.29	16.69	3.17	102.22
[1]Annual mean temperature (BIO1; °C) γ	–	4.45	3.51	–5.00	16.00
[1]Precipitation of coldest quarter (BIO19; cm)	–	225.84	177.18	20.00	1290.00
[1]Mean annual precipitation (BIO12; °C)	–	813.72	385.46	197.56	3252.94
[1]Maximum temperature of warmest month (BIO5; °C)	–	24.30	3.75	12.85	34.32
[1]Minimum temperature of coldest month (BIO6; °C)	–	–11.96	4.24	–20.60	0.44
[1]Temperature annual range (BIO7; °C)	–	36.26	3.68	22.37	47.42
[1]Mean temperature of coldest quarter (BIO11; °C)	–	–5.46	3.52	–13.00	5.00
[1]Precipitation of wettest month (BIO13; cm)	–	111.87	64.31	28.50	494.39
[1]Frost days (days)	–	216.96	53.07	55.11	338.22
[1]Humidity (Pa)	–	544.81	172.30	299.48	1354.89
[1]Annual maximum temperature (°C)	–	11.47	3.53	2.18	22.23
[1]Annual minimum temperature (°C)	–	–2.53	3.73	–11.37	9.25

Notes: SD = standard deviation. BIO = the "bioclim" variable (Nix 1986; www.worldclim.org/bioclim.htm) that we calculated via ARC AML script (Hijmans 2006), using the Daymet climate dataset.
§ Included in Maxent and GARP. γ Included in logistic regression. ¥ Included in CART.
Dashes indicate that the variable was excluded from the Maxent model due to multicollinearity. na = not applicable.

Data sources.
[1]Daymet: www.daymet.org/
[2]Base-flow index: http://water.usgs.gov/lookup/getspatial?bfi48grd
[3]Geology: http://pubs.usgs.gov/dds/dds11/
[4]National Elevation Dataset: http://ned.usgs.gov/
[5]MODIS Vegetation Indices: http://edcdaac.usgs.gov/modis/dataproducts.asp#mod13
[6]National Land Cover Dataset (2001): www.mrlc.gov/mrlc2k_nlcd.asp

References

Hijmans RJ. 2006. MkBCvarsAML version 2.3. www.worldclim.org/mkBCvars.aml. Viewed 29 Oct 2008.

Nix HA. 1986. A biogeographic analysis of Australian elapid snakes. In: Longmore R (Ed). Australian flora and fauna series 8. Canberra, Australia: Australian Government Publishing Service.