How many variables does Wordclim have, really? Generative A.I. unravels the intrinsic dimension of bioclimatic variables.

Russell Dinnage[1]

[1]Institute of Environment, Department of Biological Sciences, Florida International University

June 12, 2023

Abstract

The 19 standard bioclimatic variables available from the Worldclim dataset are some of the most used data in ecology and organismal biology. It is well know that many of the variables are correlated with each other, suggesting there is less than 19 independent dimensions of information in them. But how much information is there? Here I explore the 19 Worldclim bioclimatic variables from the perspective of the manifold hypothesis: that many high dimensional datasets are actually confined to a lower dimensional manifold embedded in an ambient space. Using a state of the art generative probabilistic model (variational autoencoder) to model the data on a non-linear manifold reveals that only 5 uncorrelated dimensions are adequate to capture the full range of variation in the bioclimatic variables. I show that these 5 variables have meaningful structure and are sufficient to produce species distribution models (SDMs) nearly as good and in some ways better than SDMs using the original 19 bioclimatic variables. I have made the 5 synthetic variables available as a raster dataset at 2.5 minute resolution in an R package that also includes functions to convert back and forth between the 5 variables and the original 19 (https://github.com/rdinnager/biocman).

Introduction

The WorldClim dataset, often referred to as bioclim, is a widely utilized resource providing detailed global climate data suitable for a variety of ecological and biological applications (Fick and Hijmans, 2017). The dataset comprises 19 bioclimatic variables, each of which represents a different aspect of annual trends, seasonality, and extreme or limiting environmental factors. These variables are derived from monthly temperature and rainfall values, providing a more biologically meaningful interpretation of environmental conditions. The set of 19 variables were selected over the period 1984-1996 as part of the development of BIOCLIM, the first species distribution modelling package (Booth et al., 2014; Booth, 2022).

The use of bioclim variables in species distribution modeling is well documented in the literature. These models frequently employ bioclim data to establish the ecological niche of a species, linking occurrences to the environmental conditions under which they are found (Elith and Leathwick, 2009). One reason the dataset is so widely used, besides its easy and wide accessibility, is that equivalent variables have been estimated for past and future climates based on paleoclimatic reconstructions and future climate model projections respectively. This makes it straight-forward to use species distribution models for prediction into the past and future, especially to predict shifts in species distributions under climate change scenarios, providing important information for conservation planning (Pearson et al., 2006).

However, a critical consideration when using BioClim variables is their inherent multicollinearity, due to the shared underlying data used in their computation. Techniques such as Principal Component Analysis

(PCA), Variance Inflation Factor (VIF) analysis, and Regularization methods have been employed to address this issue, with the aim of producing more reliable and interpretable model outcomes (Dormann et al., 2013; Kriticos et al., 2014).

More generally, though each of the 19 bioclimatic variables measure an independently meaningful and potentially useful aspect of climate, when put together they likely contain considerable redundancy of information. Another way of putting this is that the 'true' or intrinsic dimension of the data is less that the total number of dimensions in the data. One way to deal with this is using Principle Components Analysis (PCA: Pearson 1901) to reduce the dimensionality of the dataset, a frequently used strategy when using bioclim data. Classic PCA based dimension reduction is in fact a simple method of learning a data manifold, but specifically a linear one. Using PCA on bioclim can reduce the first 95% of variance into 10 orthogonal variables, but a closer look at the PCA results suggests that a linear dimension reduction method is inadequate (e.g. the data is not well confined to a hyperplane). This is apparent from the fact that the cumulative variance explained by each additional dimension forms a smooth relationship (Figure 1), whereas we expect a discontinuous 'jump' or 'elbow' in this plot in the case of a true linear manifold (Cattell, 1966).

Manifold learning is an emerging concept rooted in the manifold hypothesis, which postulates that high-dimensional real-world data lie on or near a lower-dimensional manifold embedded within the high-dimensional space (Bengio et al., 2012). This hypothesis forms the basis for a wide variety of advanced dimension reduction and data analysis techniques, which include the new and exciting class of models know as deep generative learning. Here I explore the use of one deep generative model known as a variational autoencoder (VAE: Kingma and Welling 2013) for analysing bioclim variables, a simple but nevertheless useful example which I hope also demonstrates the exciting greater potential of VAEs and other generative models for ecology.

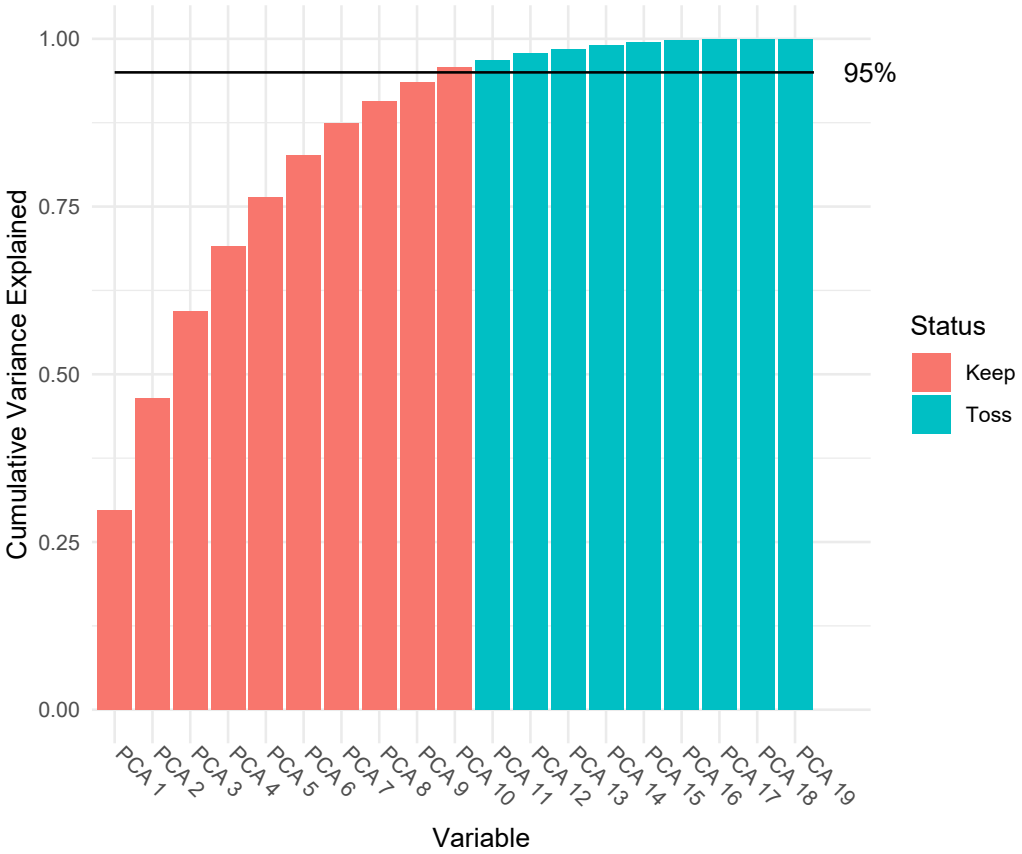

Figure 1: Cumulative variance explained by 19 PCA dimensions of Bioclim variables. There is no obvious jump in the amount of new variance explained from one PCA dimension to the next, which would indicate a possible linear manifold. Otherwise, convention suggests arbitrarily keeping the PCA dimensions that cumulatively sum up to >95% of variance explained.

Methods

Variational Autoencoders

Variational autoencoders (VAEs) are a class of generative models that employ deep learning techniques for manifold learning (Kingma and Welling, 2013). VAEs offer a probabilistic approach for representing complex, high-dimensional data in a lower-dimensional latent space. These models are built on the principles of autoencoders, which are neural networks designed to learn efficient codings of input data (Hinton and Salakhutdinov, 2006). However, VAEs introduce a layer of probabilistic inference, framing the encoding and decoding process as a probabilistic graphical model and utilizing Bayesian variational inference for learning and sampling (Kingma and Welling, 2013).

Notably, VAEs impose a specific structure on the latent space: they encourage the latent variables to follow a unit Gaussian distribution, which promotes a continuous and well-structured latent space that facilitates the learning of manifold structures (Doersch, 2016). This property is particularly useful in manifold learning tasks, where the goal is to uncover the underlying manifold structure of complex high-dimensional data. Dai and Wipf (2019) and Zheng et al. (2023) recently showed that manifold dimensions can be identified by latent variables who's posterior distributions have a low variance.

VAEs have begun to receive attention in molecular biology for a variety of uses such as visualisation (Battey et al., 2020), probability density estimation in high dimension (Ziegler et al., 2023), generating candidate functional proteins (Hawkins-Hooker et al., 2020), and discovering interpretable representations for downstream tasks (Gomari et al., 2022), but they have yet to widely used in ecology. I hope this study, in addition to providing interesting new insights to a classical dataset used in ecology, can act as a concise introduction to VAEs for ecologists, and to promote their wider use, as their potential for ecology is substantial.

VAEs work by estimating two key functions, which are approximated using neural networks: an encoder, and a decoder.

The encoder, also known as the recognition or inference model, takes the input data and provides a probabilistic encoding in the latent space. Let's say we have a dataset $X = x_1, x_2, ..., x_n$ and we wish to encode it into a latent variable z. The encoder would provide us with a distribution $q_\phi(z|x)$ over the latent variable, given the input data.

The decoder, also known as the generative or the likelihood model, takes a point in the latent space and generates a sample in the data space. It provides us with a distribution $p_\theta(x|z)$ over the data given a latent variable.

The objective of VAE is to maximize the evidence lower bound (ELBO) on the log-likelihood of the data:

$$\log p_\theta(x) \geq \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - D_{KL}(q_\phi(z|x)||p_\theta(z))$$

The first term on the right-hand side represents the expected reconstruction error — it encourages the VAE to reconstruct the input data accurately. The second term represents the Kullback-Leibler (KL) divergence between the approximate posterior $q_\phi(z|x)$ and the prior $p_\theta(z)$, which is typically chosen to be a standard normal distribution. This term acts as a regularization, encouraging the distribution of the latent variables to stay close to a unit Gaussian.

Training a VAE involves finding the parameters θ and ϕ that maximize the ELBO using stochastic gradient ascent. The "reparameterization trick" is used to enable backpropagation of the gradient through the random sampling operation during training. After training, new data can be generated by sampling from the prior $p_\theta(z)$, then passing these samples through the decoder.

A particularly useful aspect of VAEs for manifold learning is that maximizing the ELBO encourages the model to converge on a minimal set of 'active' latent variables to explain the data. A VAE learns an approximate Bayesian posterior distribution for each of the latent variables. Variables that contribute strongly to the reconstruction of the data tend to have low variance. On the other hand, any variables that are not important for reconstructing data will tend to have their variances converge to 1, consistent with the prior distribution, and facilitated by the flexibility of the decoder function, which can learn to ignore these high variance dimensions. Therefore the number of active dimensions can be deduced by the number of variables with low variance posteriors (close to zero). Recent work has shown this to be a good estimate of the true manifold dimension of simulated data (Dai and Wipf, 2019; Zheng et al., 2023).

The encoder and decoder neural networks were generated using the in-development dagnn R package (`https://github.com/rdinnager/dagnn`), which is designed to make specifying complex neural network architecture simple by specifying a directed acyclic graph (DAG) using R formula syntax.

Model Training

I trained a VAE model coded in R with the torch package (Falbel and Luraschi, 2022) on 19 global bioclimatic variables at 2.5 minute resolution, for a total of 12,532,612 data points, one for each global grid-cell. Full code to reproduce the model is available at: `https://github.com/rdinnager/bioclim_intrinsic_dimension`. I allowed for a maximum of 64 latent dimensions. The bioclim variables were standardised to zero mean

and unit standard deviation before model input. The VAE was trained using the procedure recommended in Dai and Wipf (2019), which includes as trainable 'gamma' parameter. 'Gamma' in this case is the variance of the Gaussian distribution used to calculated the log likelihood of the VAE reconstruction. This gamma parameter tends to decrease during training as the reconstruction gets better, which has the effect of increasing the weighting of the reconstruction loss relative to the KL loss as the training goes on. Dai and Wipf (2019) showed this leads to a much better reconstruction whilst still maintaining a well-structured latent space.

Interpretation

To explore how to interpret the manifold variables extracted by the VAE, I looked at how the they were related to the original 19 variables, as well as how they were represented within the world's major biomes, using data from Dinerstein et al. (2017).

Comparing Species Distribution Models

I ran three sets of species distribution models (SDMs) using the *ENMTools* R package (Warren et al., 2021): 1) SDMs using all 19 original bioclim variables; 2) SDMs using the first 10 bioclim PCA axes; and 3) SDMs using the 5 manifold variables estimated in this study. This was done using three different SDM methods: Random Forest, GLM, and the Bioclim model. The three sets of models were run on occurrence data taken from Elith et al. (2020), which is available in the R package *disdat.* I used 5 out of the 6 datasets provided by Elith et al. (2020) (the 6th proved too large to use easily in the parallelized pipeline developed for this study). The 5 datasets contained a total of 206 species from 5 different areas:

1) Australian Wet Tropics (Queensland, Australia): 20 bird species, and 20 plant species

2) Ontario, Canada: 30 bird species

3) New South Wales, Australia: 7 bat species, 10 bird species, 29 plant species, 8 small reptile species

4) New Zealand: 52 vascular plant species

5) Switzerland: 30 tree species

For each species and model combination I ran 10 replicate SDMs using ENMTools, holding out a random 10% of the occurrence points in each replicate. I then calculated the Reciever-Operator Curve Area Under the Curve metric (AUC) of model fit on model predictions for the held out 10%. I also calculated the True Skill Statistic (TSS) as an alternative measure of model fit. Additionally, I used the *ENMTools*' `env.evaluate()` function to calculate AUC and TSS in environmental space. This uses latin hypercube sampling in environmental space to sample background points which can be used in the same way that geographic background points are used to estimate the ability of the model to predict presence or absence in environmental space.

To compare the bioclim manifold variables and the bioclim PCA variables I calculated a metric ΔAUC for each replicate of each species and model, which is the different between the AUC metric of the focal SDM (manifold or PCA) and the baseline model, which in this case is the corresponding model run on all 19 original bioclim variables.

I report $\overline{\Delta\text{AUC}}$ in the main text, which is the average ΔAUC over all replicates, models and species. In order to test whether this value was significantly different than zero, I fit a Linear Mixed Model (LMM) using the *lme4* R package, including the variable type (manifold vs. PCA) as a fixed effect and species, area (e.g. New Zealand vs. Canada), and model as random effects. The reported $\overline{\Delta\text{AUC}}$ values are the estimated marginal means from this model (calculated using the *emmeans* R package). The same was done for the True Skill Statistic.

Lastly, the whole above procedure was repeated but using a simple spatial block technique for generating the hold-out sample instead of a simple random sample. For this I used the spatial block sampling option in ENMTools model fitting function.

Results

The Instrinsic Dimension of WorldClim

I trained a VAE model coded in R with the torch package (Falbel and Luraschi, 2022) on the full set of 19 bioclimatic variables at 2.5 minute resolution, for a total of 12,532,612 data points (full code available at: `https://github.com/rdinnager/bioclim_intrinsic_dimension`). I allowed for a maximum of 64 latent dimensions. The data was standardised before model input.

After model training, examining the posterior distribution of the 64 latent dimensions revealed only 5 had low variance (Figure 2), suggesting an intrinsic dimensionality of 5. Using only these 5 variables to reconstruct the original data through the trained decoder, the error is extremely low, with a mean squared error of 0.0036 (on the standardised variables, so in units of standard deviations).

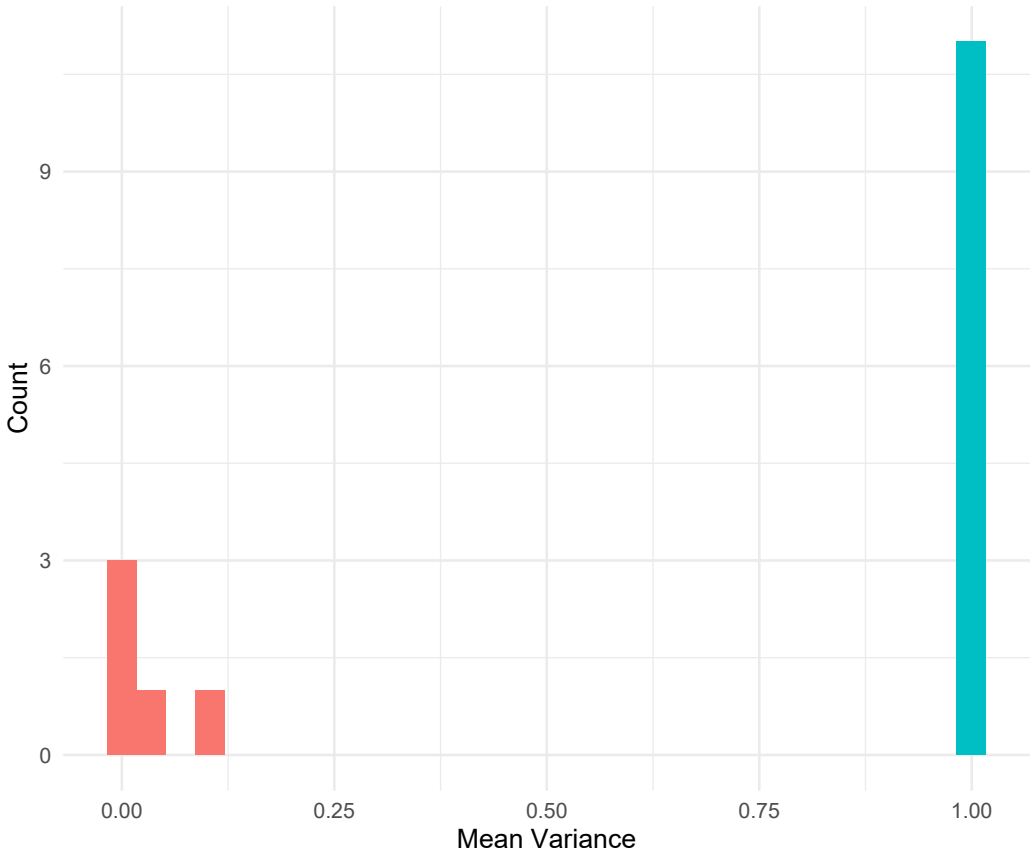

Figure 2: The mean variances along each variational autoencoder latent dimension (there were 16 altogether). All dimensions except for 5 had variance near or equal to 1, matching the prior distribution. The remaining 5 had variances close to zero. This is indicative that these are 'manifold' dimensions: they are necessary to reconstruct the full data. All other dimensions are just noise that the decoder network learns to ignore.

Using SDM benchmark data from Elith et al. (2020), and three different SDM methods (random forest, glm, and bioclim) from the ENMTools R package (Warren et al., 2021), I additionally tested whether these 5 'latent' variables are useful by comparing species distribution models fit with them with models fit on all 19 original variables, as well as the first 10 PCA dimensions as a comparison. For the 5 manifold dimensions and the PCA axes I calculated the model performance using $\overline{\Delta\text{AUC}}$, defined as the average difference in AUC between the focal model and the corresponding model using all 19 original variables, averaged over all species and methods tested. I find the 5 variables perform slightly less well (but not significantly) at predicting held-out test samples in geographic space according to AUC ($\overline{\Delta\text{AUC}}$ = -0.037 [CI: -0.076, 0.002]; -0.02 for PCA [CI: -0.058, 0.019]; Figure S6) but perform considerably better at predicting them in environmental space ($\overline{\Delta\text{AUC}}$ = +0.11 [CI: 0.071, 0.15] , +0.047 for PCA [CI: 0.0096, 0.087]; Figure S6) (Warren et al., 2018). The same analysis using the true skill statistic (TSS) instead of AUC gave similar qualitative results (Figure S8). I suspected the slightly worse performance in geographic space was because with fewer variables, models could not overfit as strongly to approximate spatial structure in the occurrence data. To test this I repeated the SDMs using a simple spatial block method for holding out test samples. This confirmed my suspicion, with the reduction in AUC for spatially held-out test data being much smaller and again, not different from zero statistically in geographic space ($\overline{\Delta\text{AUC}}$ = -0.017 [CI: -0.043, 0.0081]; Figure S7), but still much improved in environmental space ($\overline{\Delta\text{AUC}}$ = +0.097 [CI: 0.071, 0.12]; Figure S7).

Discussion

Interpreting the Manifold Dimensions

Examining the 5 active variables shows that each has a unique global pattern (Figure 3) and link to the original 19 variables (Figures S1-S5). Three of the variables admit a fairly straightforward interpretation, BIOCMAN1, BIOCMAN2, and BIOCMAN4 – their values have a mostly monotonic relationships with the original 19, and their global distribution maps show clear links to important global climate factors, BIOCMAN1 being elevation, BIOCMAN4 being latitude, and BIOCMAN4 being areas supporting rainforest, including temperate rainforest. The remaining two are less straightforward, with some relationships with both the original 19 variables being unimodal or 'U-shaped'. On the other hand, BIOCMAN5 appears to clearly differentiate areas with monsoonal climates with high values. BIOCMAN3 has the least clear interpretation but globally low values seem to be associated with arid and semi-arid ecoregions (Figure 1; Figure S3).

For the application of species distribution modelling and other ecological models where interpretation of a model's coefficients might be of interest, the issue of challenging interpretability can be mitigated using the decoder function, which allows a straightforward way to convert the manifold variables back to the original data space. In this way they are no less interpretable than using a transformation to make a variable better conform to model assumptions (e.g. log or logistic transformations to reduce covariate skew). A simple procedure to interpret model coefficients when using bioclim manifold variables as predictors is to output predictions of the model across a range of manifold variable values of interest, then plot these predictions against the original 19 variables after back-transforming the manifold variables using the decoder. This then places the model prediction back into the context of full bioclimatic variation.

Though many ecologists probably suspected that bioclim had fewer intrinsic dimensions than 19, including myself, it is perhaps surprising that the full panoply of bioclimatic variation on Earth can be reconstructed with as few as 5 variables. It is also remarkable that these 5 variables have close correspondences with known sources of climate variation such as latitude, elevation, and the monsoon. This speaks to the potential power of generative A.I. methods in ecology. In this case the physical and biological processes of the earth encode themselves in global climate variation, which this VAE was then able to decode, thanks to its ability to learn flexible non-linear functions of a simplified underlying space. A practical implication of this is we have a principled way to reduce 19 correlated variables to 5 uncorrelated that work just as well in at least one

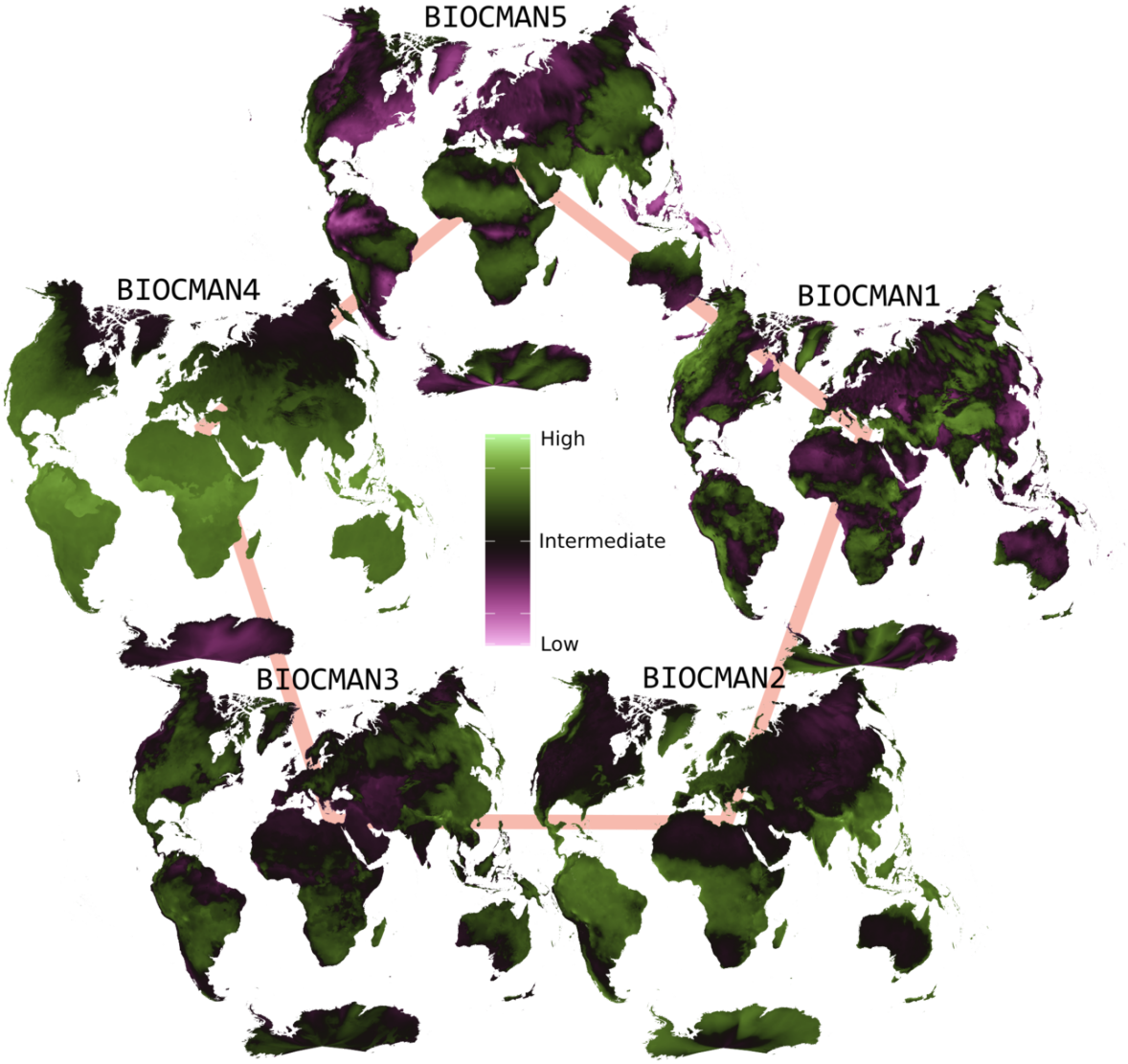

Figure 3: Global maps of the 5 VAE manifold variables. Most variables appear to be imperfect but reasonable approximations to known global climate factors. BIOCMAN1 looks like elevation. BIOCMAN2 is where rainforest exists (including temperate rainforest). BIOCMAN3 has the least clear pattern, but low values seem to be associated with arid and semi-arid places. BIOCMAN4 is close to a map of latitude, and BIOCMAN5 is high where there are monsoonal climates. Maps are presented in Danseiji V projection, which has low distortion in terrestrial areas at the cost of high distortion in the oceans (Kunimune, 2020). High resolution versions with full scale bars can be found in Figures S9-S13).

downstream application (SDM) and in at least one way they actually work better. Thanks to the structure of the model and how it encourages compact representations, this VAE uses as few dimensions as necessary to fully regenerate the original data, meaning we do not have to make arbitrary decisions on the number of dimensions to use, as is the case for most alternative dimension reduction techniques. If even just for this reason, VAE methods are currently under-used in ecology.

Supplementary Material

Viol

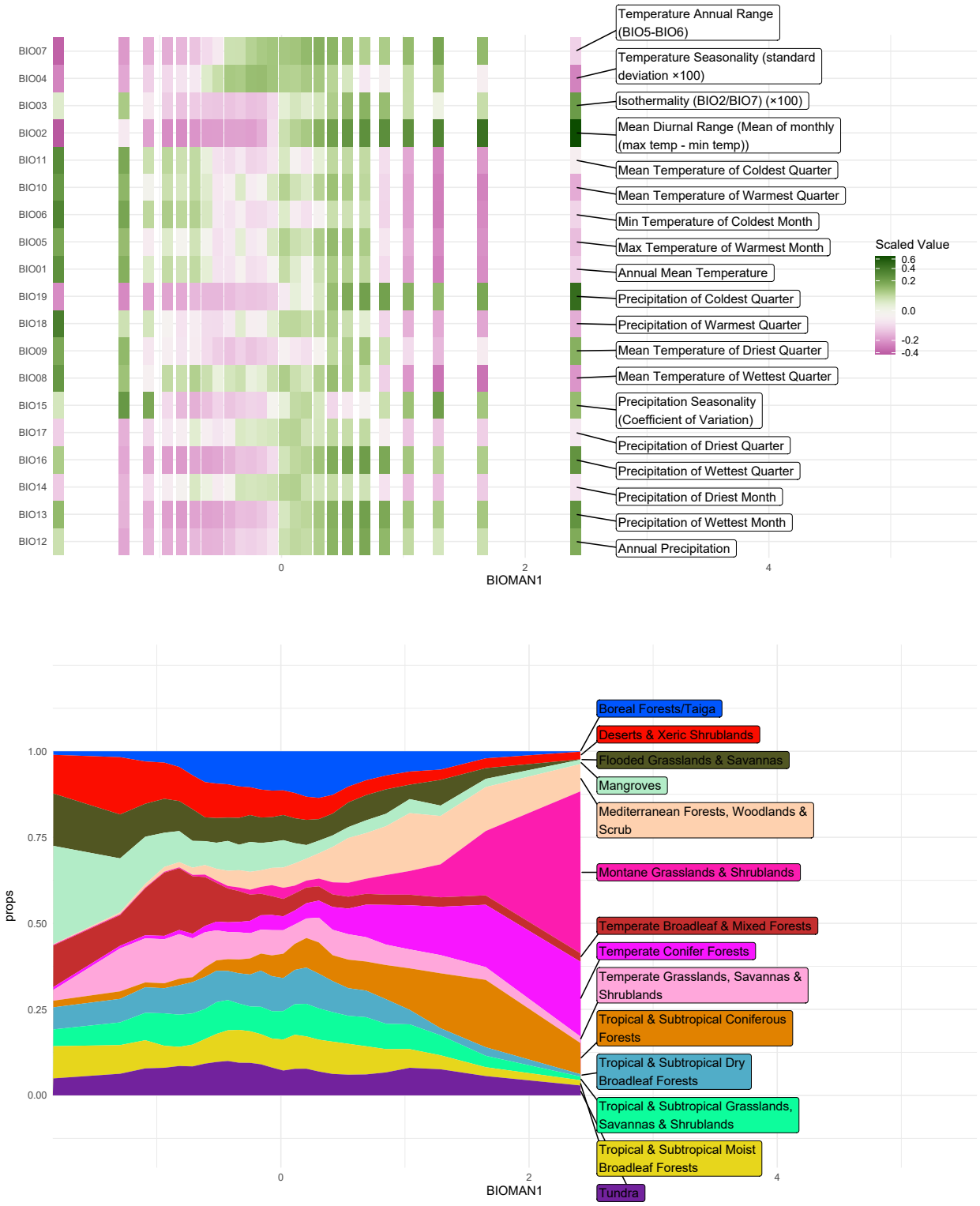

Figure S1: Examining and interpreting manifold dimension 1 of bioclim data. The top panel shows how the latent variable relates to the original 19 bioclim variable. Each row represents one of the 19 bioclim variables. Each column represents a bin of manifold variable values, where bins are calculated to have an even number of points in each bin. The coloured cells represent the mean decoded value for each of the 19 bioclim variables that correspond to the manifold variable values in each bin. This gives an idea of how the manifold variable maps onto the original variables (similar to a non-linear version of loadings in PCA). The bottom panel shows how ecoregions are represented along the manifold variable. For each manifold variable bin the proportion of points that fell in each ecoregions was calculated and then doubly standardized by the total representation of each ecoregion globally. These data were then used to make a stacked area plot, where each colour-filled area corresponds to the proportion of the ecoregion along the manifold variable gradient.

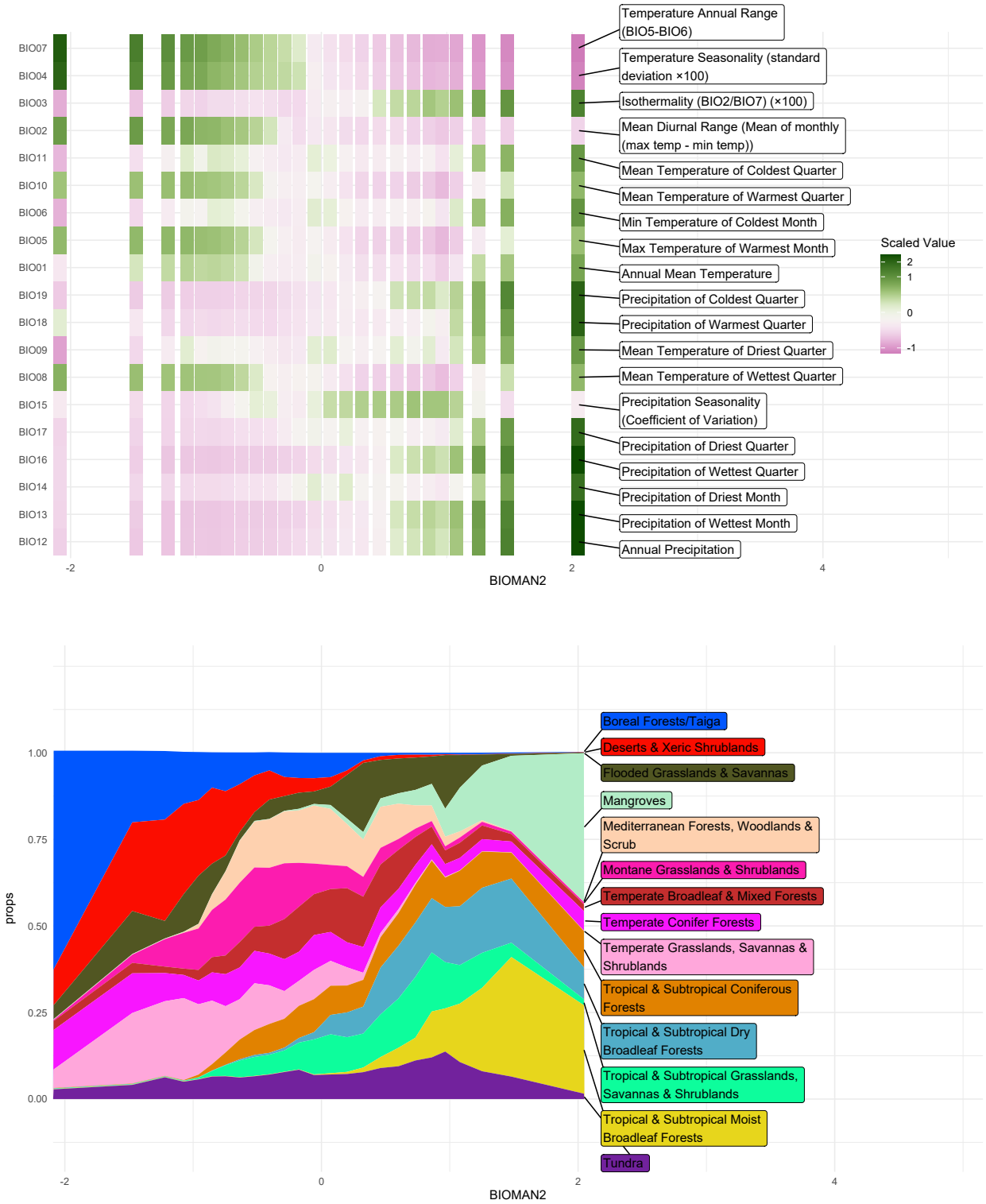

Figure S2: Examining and interpreting manifold dimension 2 of bioclim data. The top panel shows how the latent variable relates to the original 19 bioclim variable. Each row represents one of the 19 bioclim variables. Each column represents a bin of manifold variable values, where bins are calculated to have an even number of points in each bin. The coloured cells represent the mean decoded value for each of the 19 bioclim variables that correspond to the manifold variable values in each bin. This gives an idea of how the manifold variable maps onto the original variables (similar to a non-linear version of loadings in PCA). The bottom panel shows how ecoregions are represented along the manifold variable. For each manifold variable bin the proportion of points that fell in each ecoregions was calculated and then doubly standardized by the total representation of each ecoregion globally. These data were then used to make a stacked area plot, where each colour-filled area corresponds to the proportion of the ecoregion along the manifold variable gradient.

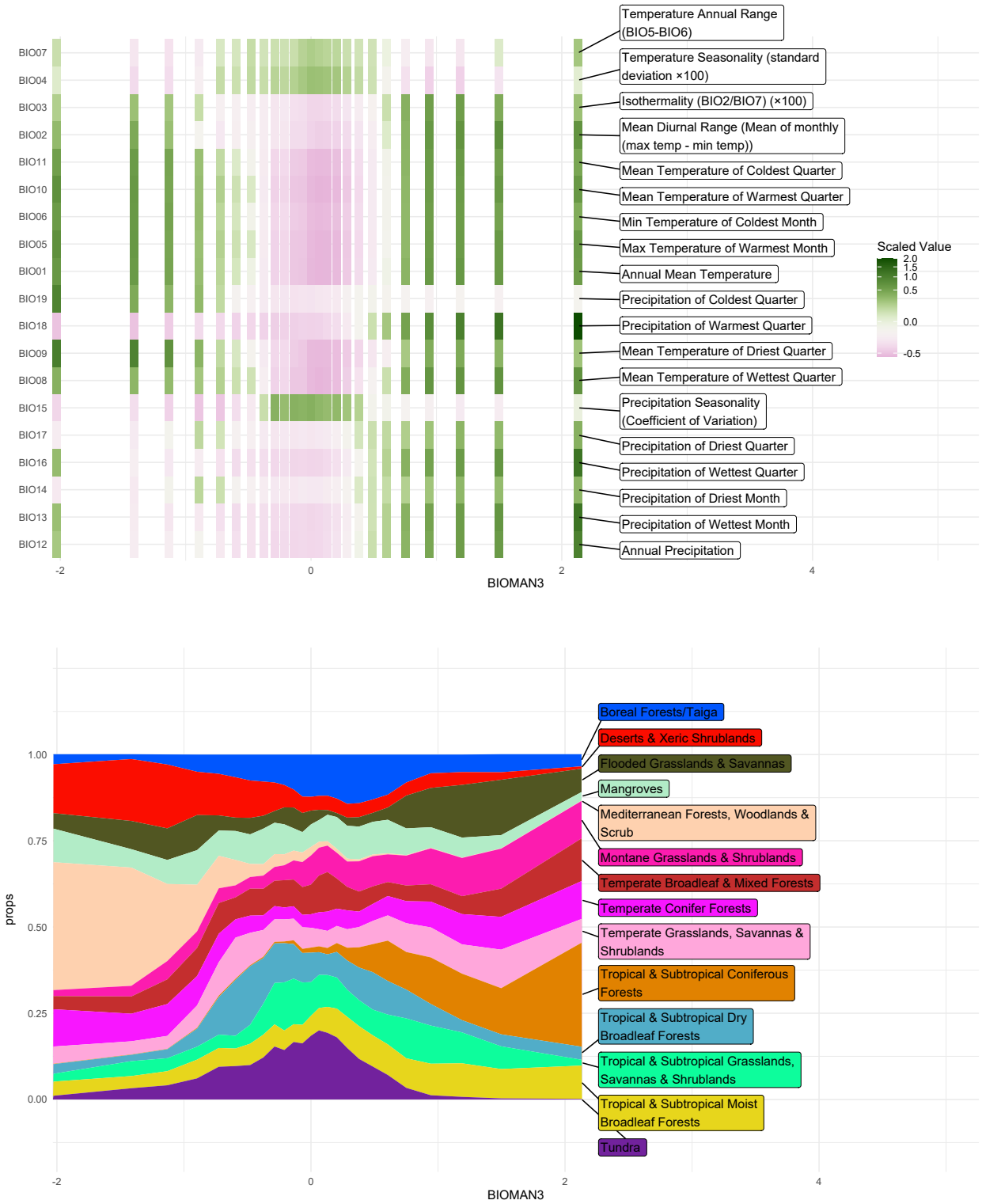

Figure S3: Examining and interpreting manifold dimension 3 of bioclim data. The top panel shows how the latent variable relates to the original 19 bioclim variable. Each row represents one of the 19 bioclim variables. Each column represents a bin of manifold variable values, where bins are calculated to have an even number of points in each bin. The coloured cells represent the mean decoded value for each of the 19 bioclim variables that correspond to the manifold variable values in each bin. This gives an idea of how the manifold variable maps onto the original variables (similar to a non-linear version of loadings in PCA). The bottom panel shows how ecoregions are represented along the manifold variable. For each manifold variable bin the proportion of points that fell in each ecoregions was calculated and then doubly standardized by the total representation of each ecoregion globally. These data were then used to make a stacked area plot, where each colour-filled area corresponds to the proportion of the ecoregion along the manifold variable gradient.

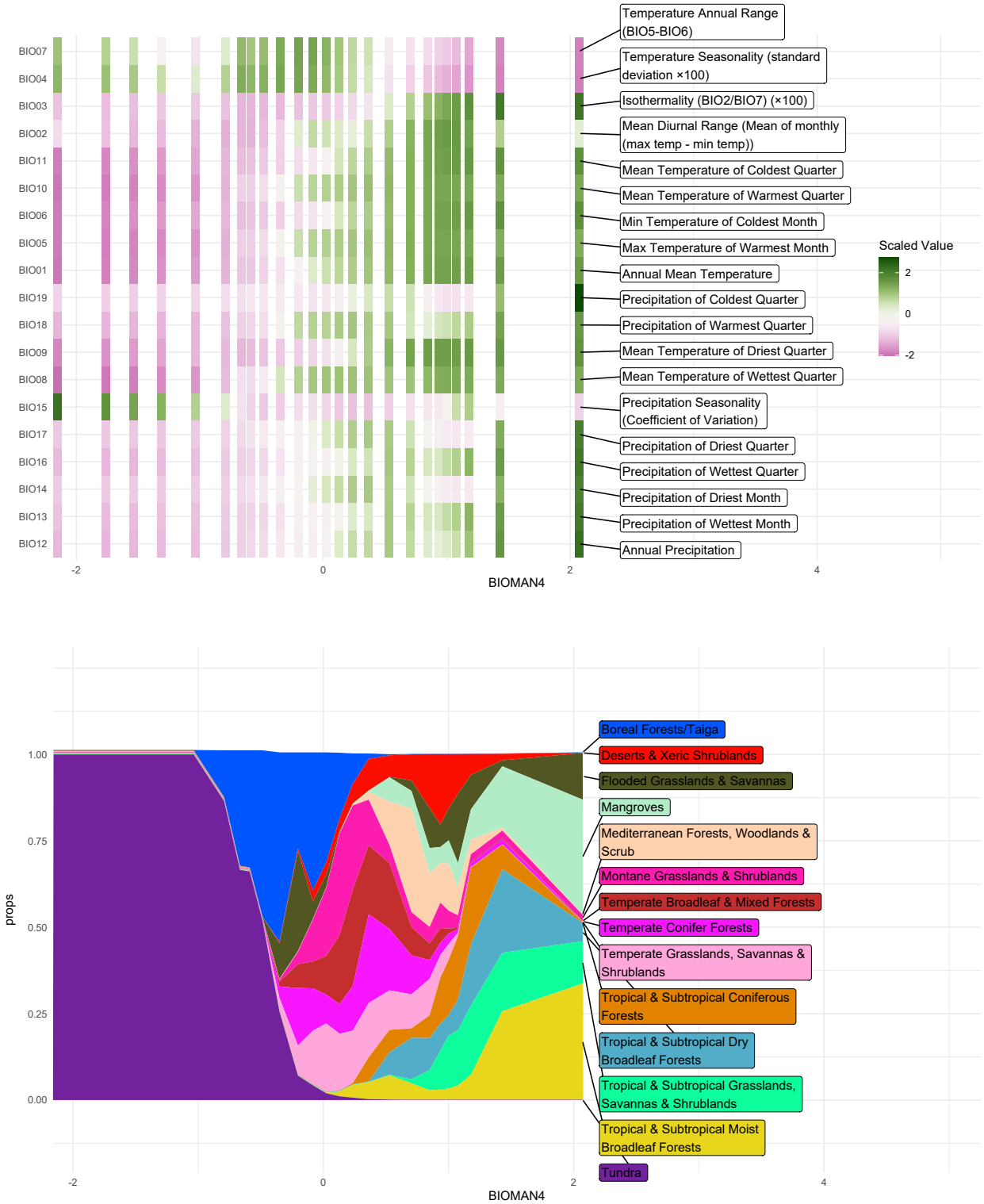

Figure S4: Examining and interpreting manifold dimension 4 of bioclim data. The top panel shows how the latent variable relates to the original 19 bioclim variable. Each row represents one of the 19 bioclim variables. Each column represents a bin of manifold variable values, where bins are calculated to have an even number of points in each bin. The coloured cells represent the mean decoded value for each of the 19 bioclim variables that correspond to the manifold variable values in each bin. This gives an idea of how the manifold variable maps onto the original variables (similar to a non-linear version of loadings in PCA). The bottom panel shows how ecoregions are represented along the manifold variable. For each manifold variable bin the proportion of points that fell in each ecoregions was calculated and then doubly standardized by the total representation of each ecoregion globally. These data were then used to make a stacked area plot, where each colour-filled area corresponds to the proportion of the ecoregion along the manifold variable gradient.

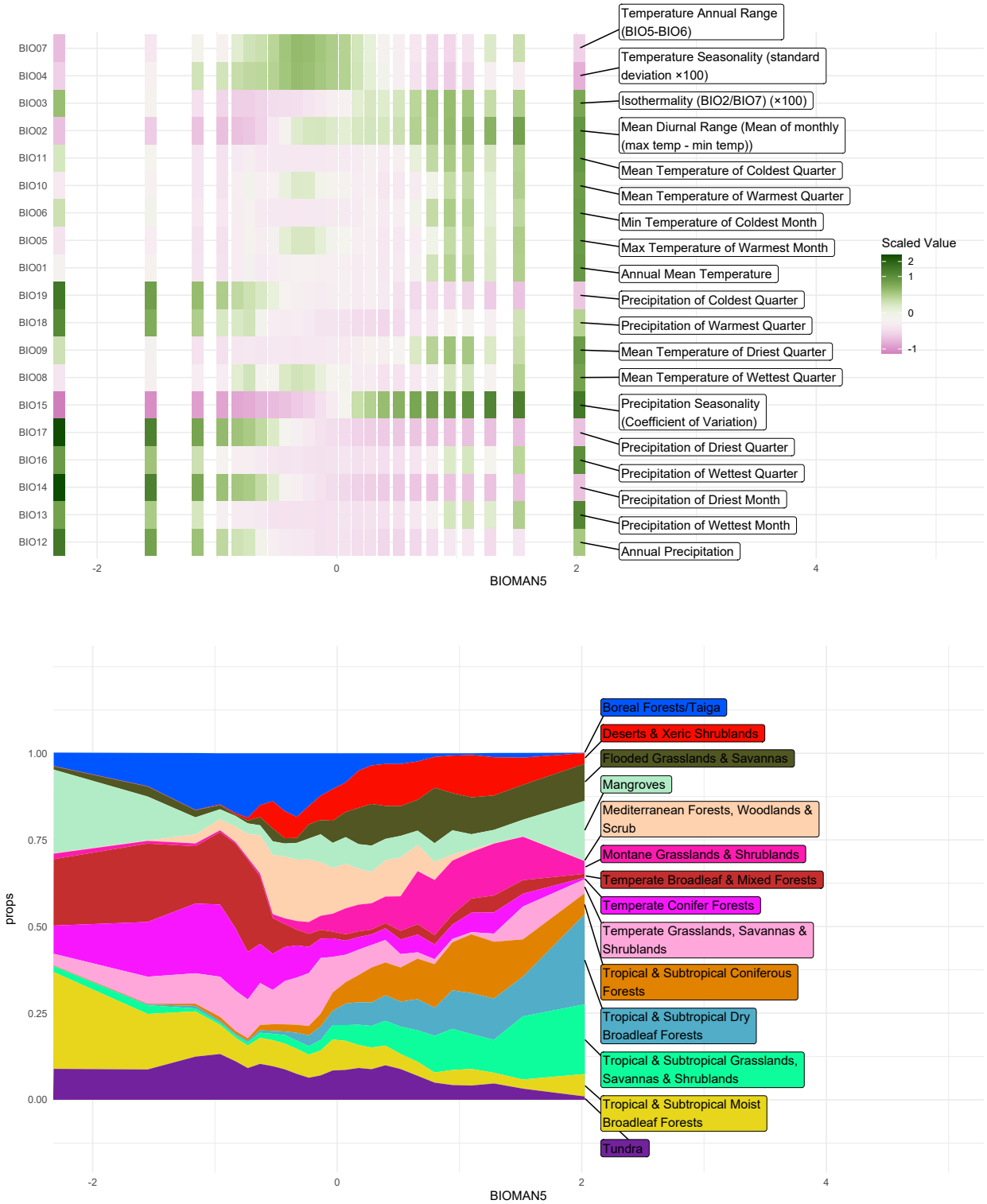

Figure S5: Examining and interpreting manifold dimension 5 of bioclim data. The top panel shows how the latent variable relates to the original 19 bioclim variable. Each row represents one of the 19 bioclim variables. Each column represents a bin of manifold variable values, where bins are calculated to have an even number of points in each bin. The coloured cells represent the mean decoded value for each of the 19 bioclim variables that correspond to the manifold variable values in each bin. This gives an idea of how the manifold variable maps onto the original variables (similar to a non-linear version of loadings in PCA). The bottom panel shows how ecoregions are represented along the manifold variable. For each manifold variable bin the proportion of points that fell in each ecoregions was calculated and then doubly standardized by the total representation of each ecoregion globally. These data were then used to make a stacked area plot, where each colour-filled area corresponds to the proportion of the ecoregion along the manifold variable gradient.

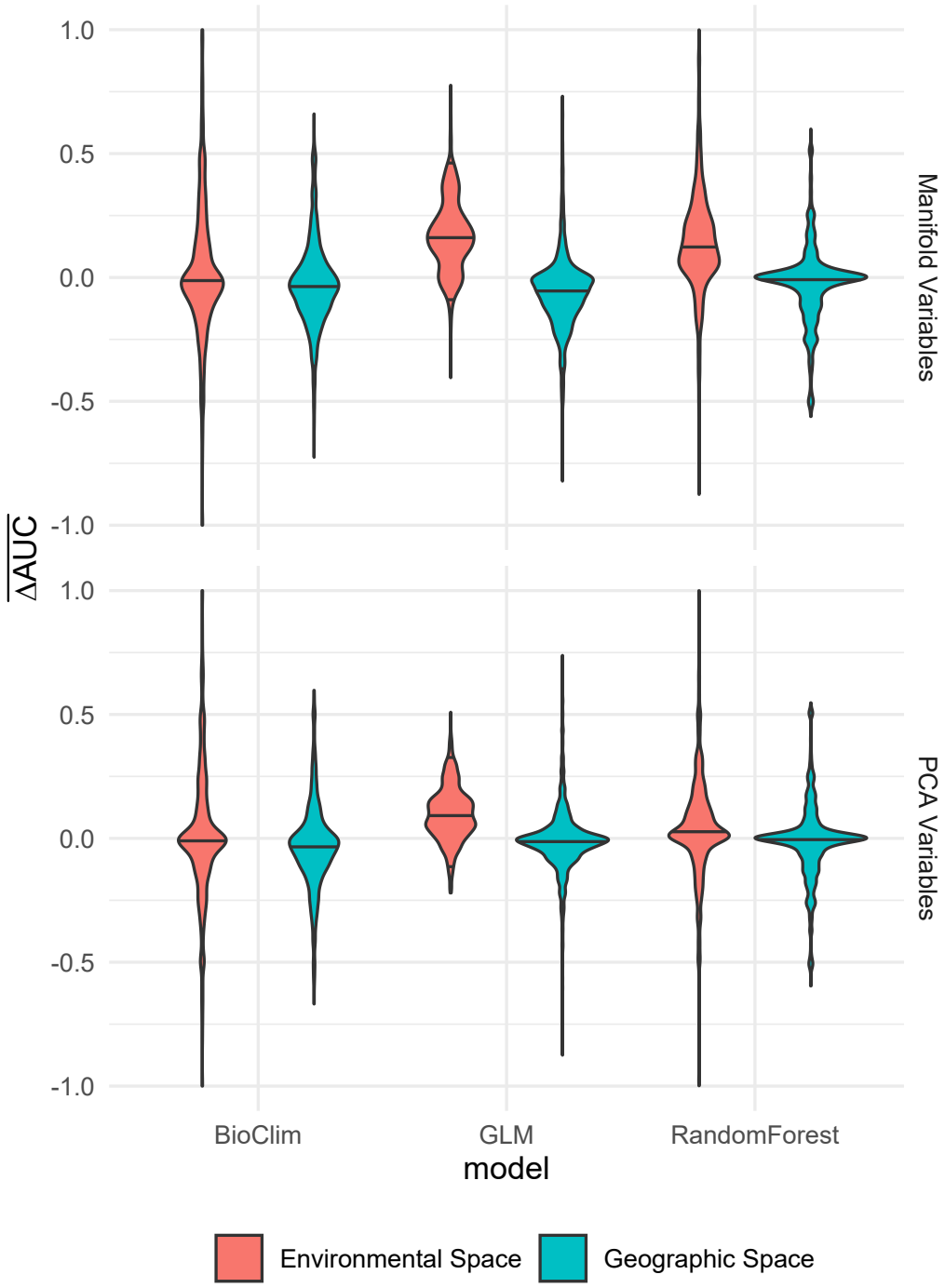

Figure S6: Violin plot of random hold-out $\overline{\Delta\text{AUC}}$, by model type (X axis), Environmental vs. Geographic space (colour) and whether the models used the manifold or the PCA variables (top and bottom panels). Width of the violins is proportional to the density of points at the value on the y axis. Horizontal black lines represent the 2.5%, 50%, and 97.5% quantiles. $\overline{\Delta\text{AUC}}$ is the difference in AUC between the focal model and the corresponding model using the original 19 bioclim variables, calculated on a 10% random hold-out sample.

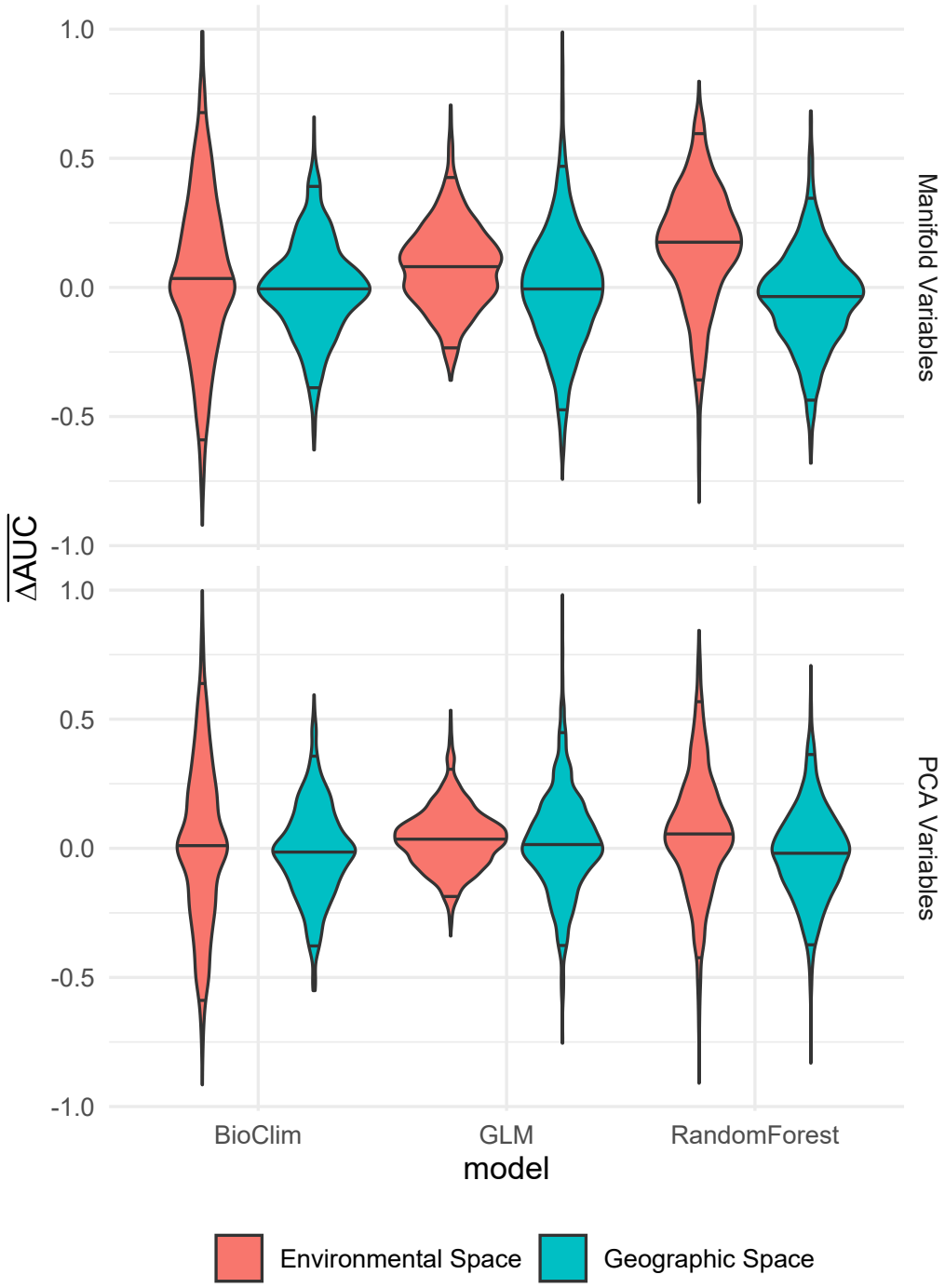

Figure S7: Violin plot of spatial block hold-out $\overline{\Delta AUC}$, by model type (X axis), Environmental vs. Geographic space (colour) and whether the models used the manifold or the PCA variables (top and bottom panels). Width of the violins is proportional to the density of points at the value on the y axis. Horizontal black lines represent the 2.5%, 50%, and 97.5% quantiles. $\overline{\Delta AUC}$ is the difference in AUC between the focal model and the corresponding model using the original 19 bioclim variables, calculated on a spatial block hold-out sample.

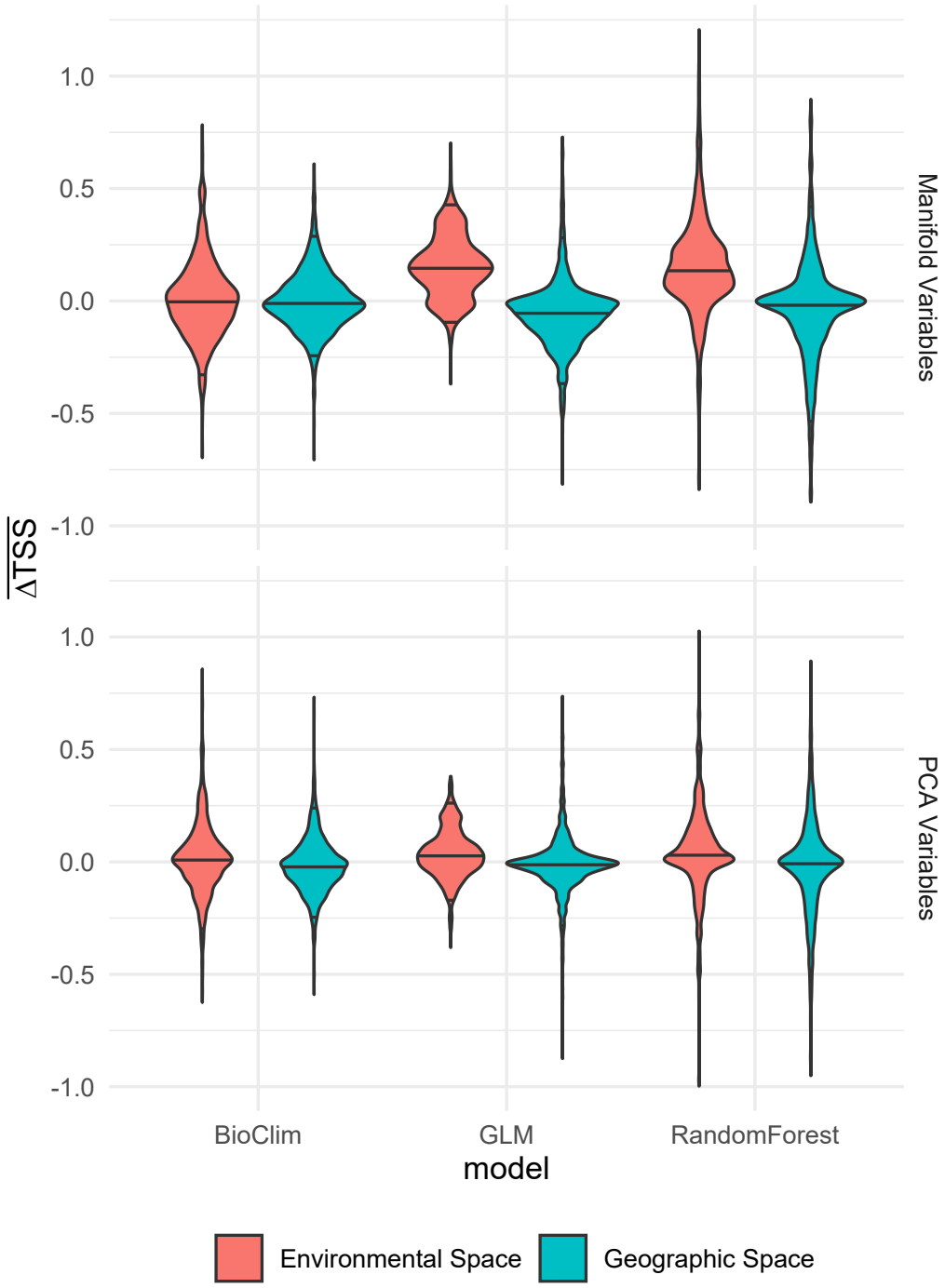

Figure S8: Violin plot of random hold-out $\overline{\Delta TSS}$, by model type (X axis), Environmental vs. Geographic space (colour) and whether the models used the manifold or the PCA variables (top and bottom panels). Width of the violins is proportional to the density of points at the value on the y axis. Horizontal black lines represent the 2.5%, 50%, and 97.5% quantiles. $\overline{\Delta TSS}$ is the difference in TSS between the focal model and the corresponding model using the original 19 bioclim variables, calculated on a 10% random hold-out sample.

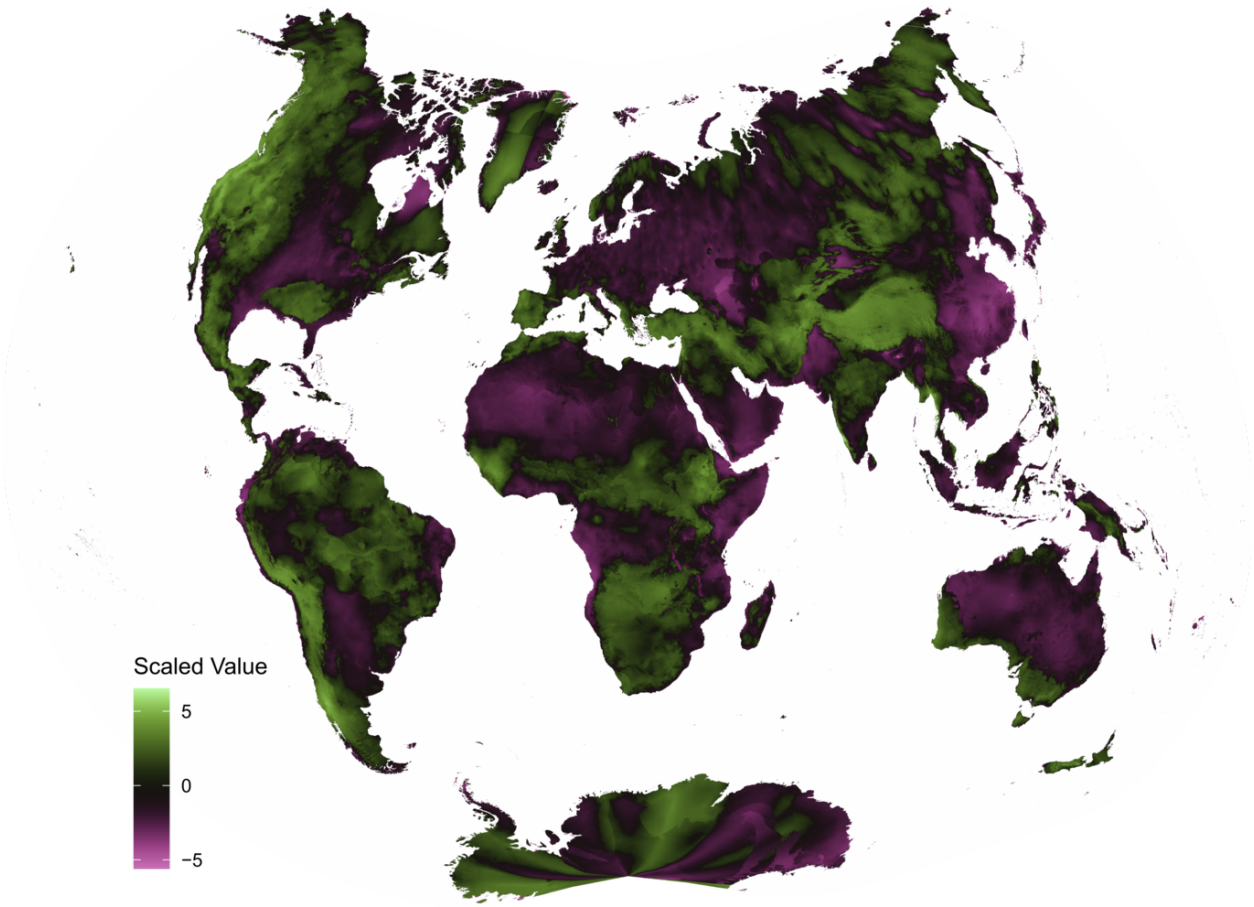

Figure S9: Map of bioclim manifold variable 1 in Daiseiji V projection (Kunimune, 2020) .

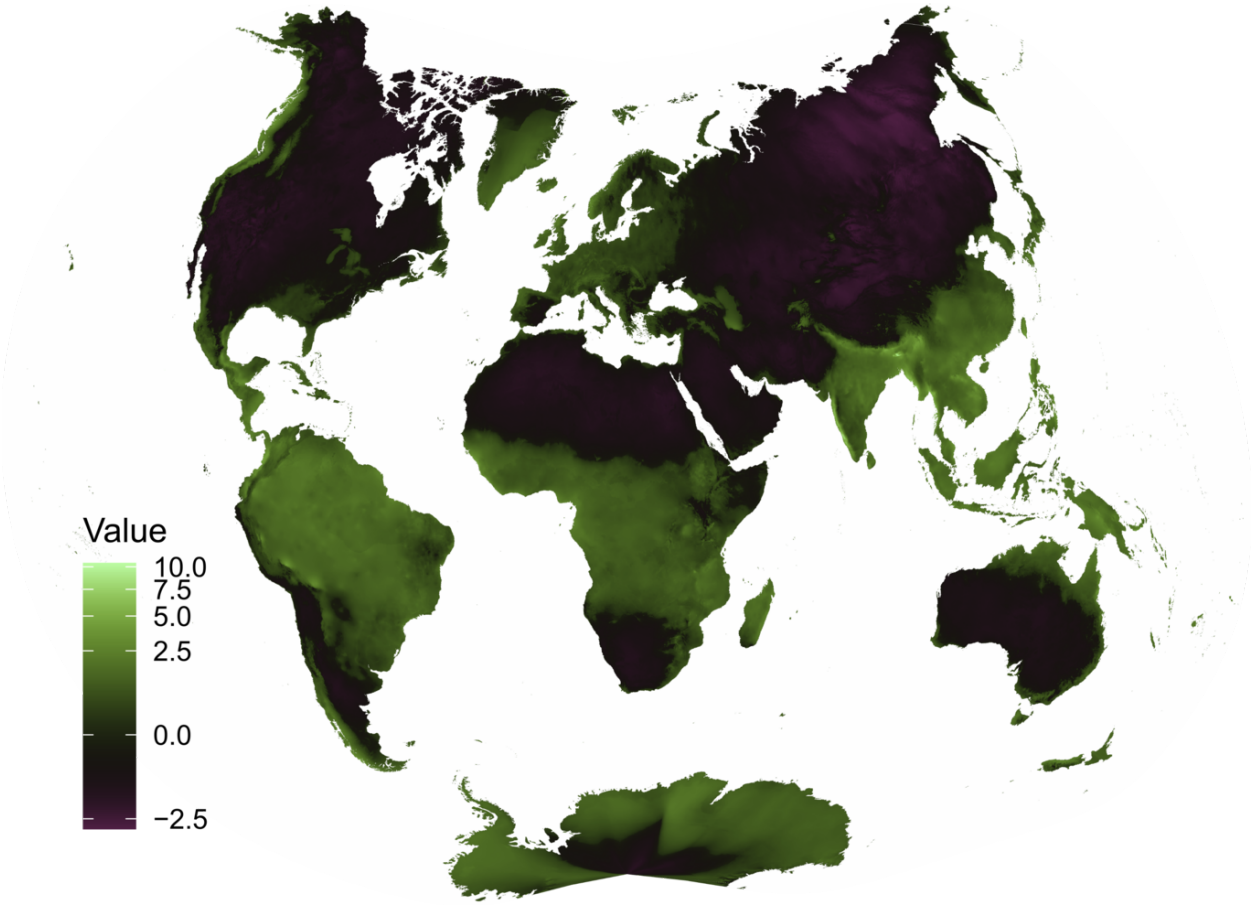

Figure S10: Map of bioclim manifold variable 2 in Daiseiji V projection (Kunimune, 2020).

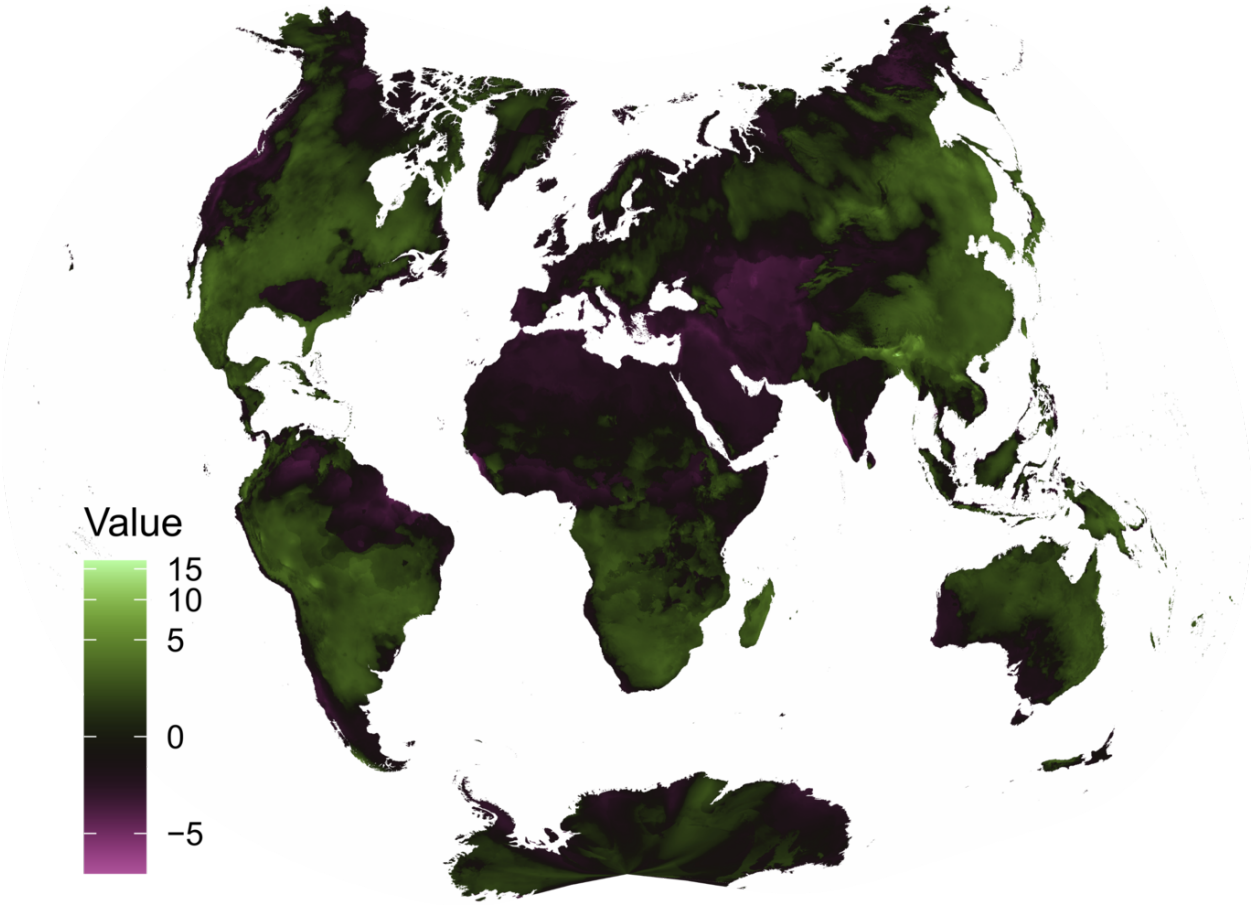

Figure S11: Map of bioclim manifold variable 3 in Daiseiji V projection (Kunimune, 2020).

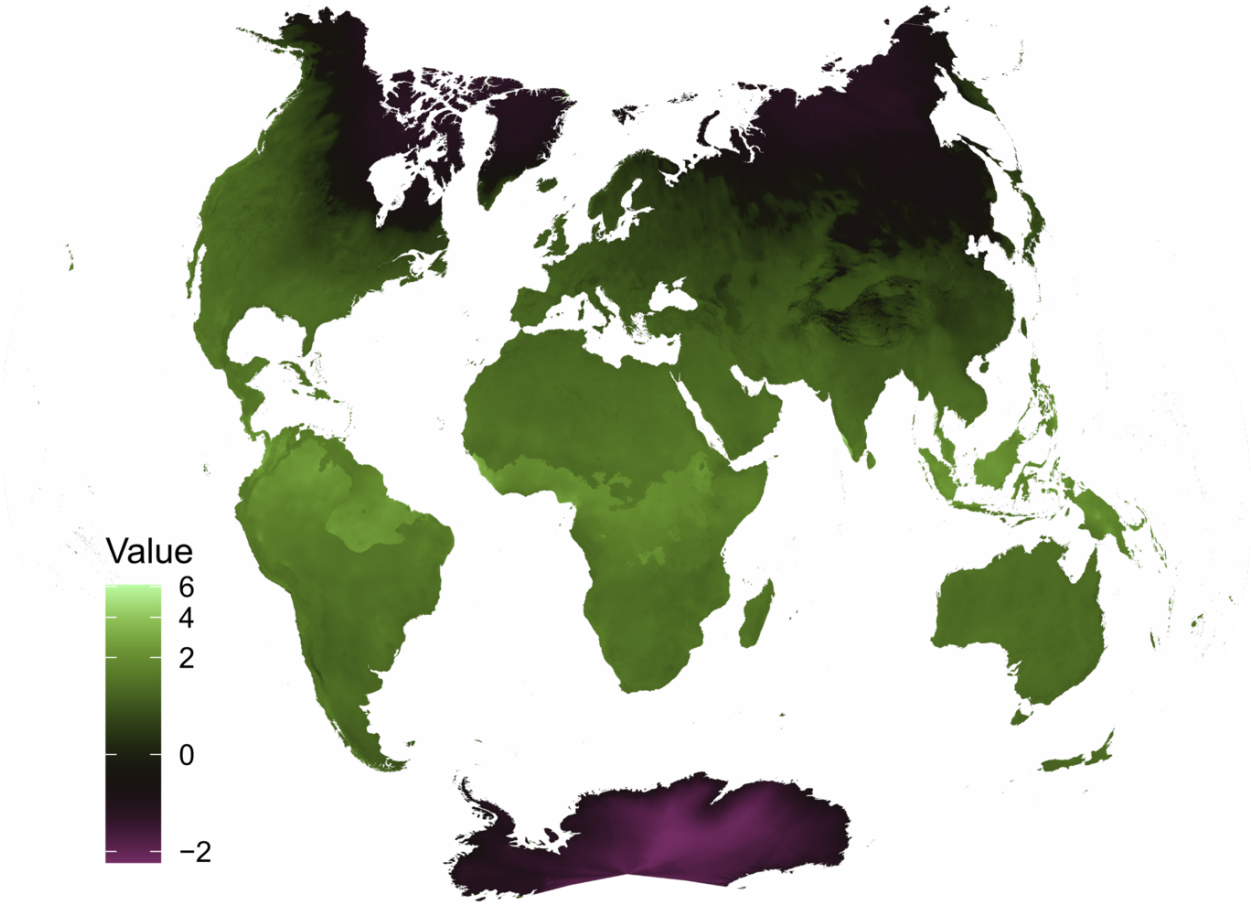

Figure S12: Map of bioclim manifold variable 4 in Daiseiji V projection (Kunimune, 2020).

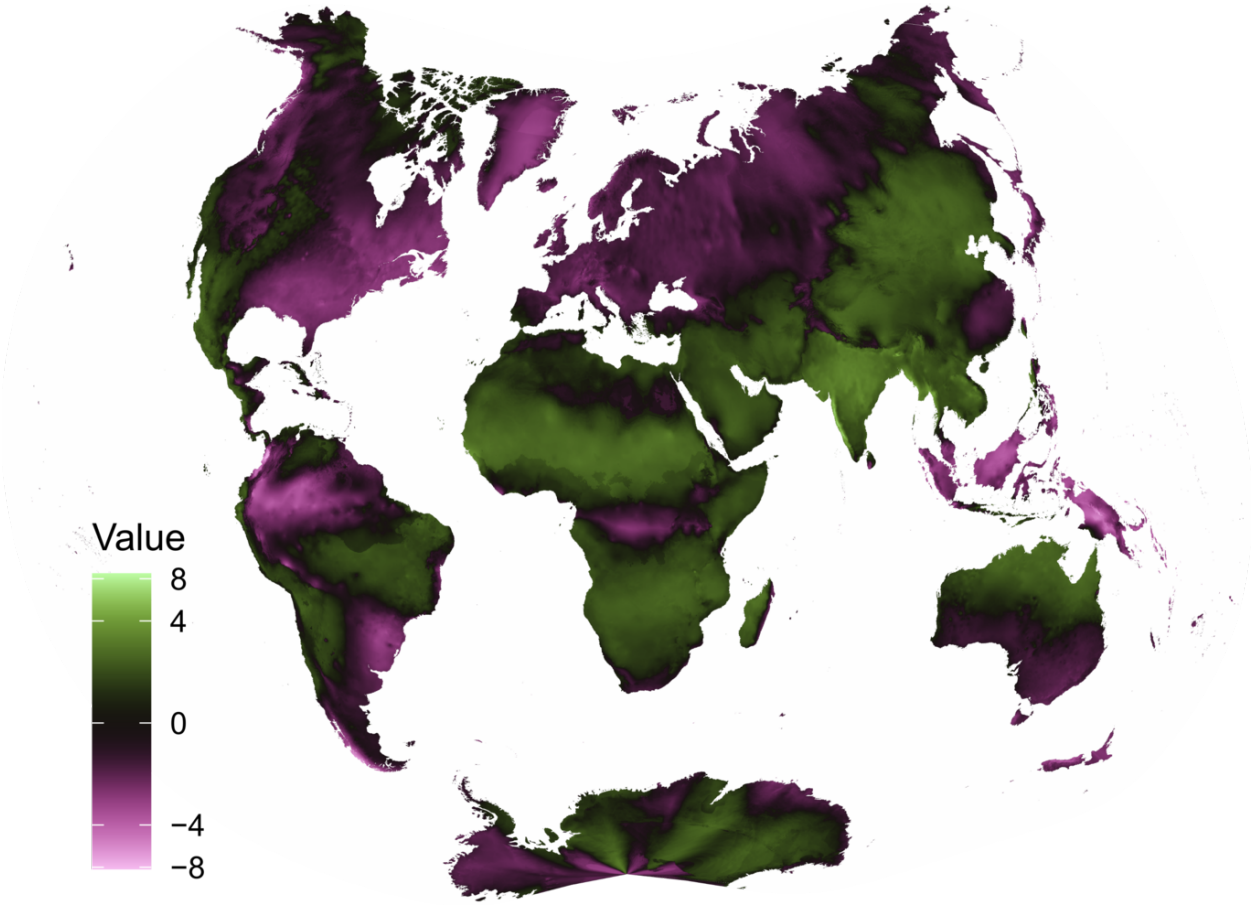

Figure S13: Map of bioclim manifold variable 5 in Daiseiji V projection (Kunimune, 2020).

References

C. J. Battey, Gabrielle C. Coffing, and Andrew D. Kern. Visualizing population structure with variational autoencoders. *G3: Genes—Genomes—Genetics*, 11, 2020.

Yoshua Bengio, Aaron C. Courville, and Pascal Vincent. Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35:1798–1828, 2012.

Trevor H. Booth. Checking bioclimatic variables that combine temperature and precipitation data before their use in species distribution models. *Austral Ecology*, 2022.

Trevor H. Booth, Henry Nix, John R. Busby, and Michael F. Hutchinson. bioclim: the first species distribution modelling package, its early applications and relevance to most current MaxEnt studies. *Diversity and Distributions*, 20, 2014.

Raymond Bernard Cattell. The Scree Test For The Number Of Factors. *Multivariate behavioral research*, 1 2:245–76, 1966.

Bin Dai and David Paul Wipf. Diagnosing and Enhancing VAE Models. *ArXiv*, abs/1903.05789, 2019.

Eric Dinerstein, David P. Olson, Anup R. Joshi, Carly Vynne, Neil D. Burgess, Eric D. Wikramanayake, Nathan R. Hahn, Suzanne Palminteri, Prashant Hedao, Reed F. Noss, Matt Hansen, Harvey Locke, Erle C. Ellis, Benjamin S Jones, Charles Victor Barber, Randy Hayes, Cyril F. Kormos, Vance G. Martin, Eileen Crist, Wes Sechrest, Lori Price, J. E. M. Baillie, Donald E. Weeden, Kieran F. Suckling, Crystal Davis, Nigel C. Sizer, Rebecca Moore, David Thau, Tanya Birch, Peter V. Potapov, Svetlana Turubanova, Alexandra Tyukavina, Nadia de Souza, Lilian Pintea, José Carlos Brito, Othman Abd ar Rahman Llewellyn, Anthony G. Miller, Annette Patzelt, Shahina A. Ghazanfar, John R. Timberlake, Heinz Klöser, Yara Shennan-Farpón, Roeland Kindt, Jens-Peter Barnekow Lillesø, Paulo van Breugel, Lars Graudal, Maianna Voge, Khalaf F. Al-Shammari, and Muhammad Saleem. An Ecoregion-Based Approach to Protecting Half the Terrestrial Realm. *Bioscience*, 67:534 – 545, 2017.

Carl Doersch. Tutorial on Variational Autoencoders. *ArXiv*, abs/1606.05908, 2016.

Carsten F. Dormann, Jane Elith, Sven Bacher, Carsten M. Buchmann, Gudrun Carl, Gabriel Carré, Jaime Ricardo García Márquez, Bernd Gruber, Bruno Lafourcade, Pedro J. Leitão, Tamara Münkemüller, Colin J. McClean, Patrick E. Osborne, Björn Reineking, Boris Schröder, Andrew K. Skidmore, Damaris Zurell, and Sven Lautenbach. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, 36:27–46, 2013.

Jane Elith and John R. Leathwick. Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. *Annual Review of Ecology, Evolution, and Systematics*, 40:677–697, 2009.

Jane Elith, Catherine H. Graham, Roozbeh Valavi, Meinrad Abegg, Caroline Bruce, Andrew James Ford, Antoine Guisan, Robert J. Hijmans, Falk Huettmann, Lúcia Garcez Lohmann, Bette A. Loiselle, Craig C Moritz, Jacob McC. Overton, Andrew Townsend Peterson, Steven J. Phillips, Karen S. Richardson, Stephen E. Williams, Susan K. Wiser, Thomas Wohlgemuth, and Niklaus E. Zimmermann. Presence-only and Presence-absence Data for Comparing Species Distribution Modeling Methods. *Biodiversity Informatics*, 2020.

Daniel Falbel and Javier Luraschi. *torch: Tensors and Neural Networks with 'GPU' Acceleration*, 2022. URL https://CRAN.R-project.org/package=torch. R package version 0.9.0.

Stephen E. Fick and Robert J. Hijmans. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, 37, 2017.

Daniel P. Gomari, Annalise Schweickart, Leandro Cerchietti, Elisabeth Paietta, Hugo Fernandez, Hassen Al-Amin, Karsten Suhre, and Jan Krumsiek. Variational autoencoders learn transferrable representations of metabolomics data. *Communications Biology*, 5, 2022.

Alex Hawkins-Hooker, Florence Depardieu, Sebastien Baur, Guillaume Couairon, Arthur Chen, and David Bikard. Generating functional protein variants with variational autoencoders. *PLoS Computational Biology*, 17, 2020.

Geoffrey E. Hinton and Ruslan Salakhutdinov. Reducing the Dimensionality of Data with Neural Networks. *Science*, 313:504 – 507, 2006.

Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. *CoRR*, abs/1312.6114, 2013.

Darren J. Kriticos, Vojtěch Jaroık, and Noboru Ota. Extending the suite of bioclim variables: a proposed registry system and case study using principal components analysis. *Methods in Ecology and Evolution*, 5, 2014.

J H Kunimune. Minimum-error world map projections defined by polydimensional meshes. *International Journal of Cartography*, 7:78 – 99, 2020.

Karl Pearson. LIII. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine Series 1*, 2:559–572, 1901.

Richard G. Pearson, Wilfried Thuiller, Miguel Bastos Araújo, Enrique Martínez-Meyer, Lluís Brotóns, Colin J. McClean, Lera Miles, Pedro Segurado, Terence P. Dawson, and David C. Lees. Model-based uncertainty in species range prediction. *Journal of Biogeography*, 33, 2006.

Dan L. Warren, Linda J. Beaumont, Russell Dinnage, and John B. Baumgartner. New methods for measuring ENM breadth and overlap in environmental space. *Ecography*, 2018.

Dan L. Warren, Nicholas J. Matzke, Marcel Cardillo, John B. Baumgartner, Linda J. Beaumont, Michael Turelli, Richard E. Glor, Nicholas A. Huron, Marianna V. P. Simões, Teresa L. Iglesias, Julien C. Piquet, and Russell Dinnage. ENMTools 1.0: an R package for comparative ecological biogeography. *Ecography*, 2021.

Yijia Zheng, Tong He, Yixuan Qiu, and David Paul Wipf. Learning Manifold Dimensions with Conditional Variational Autoencoders. *ArXiv*, abs/2302.11756, 2023.

Cheyenne Ziegler, Jonathan Martin, Claude Sinner, and Faruck Morcos. Latent generative landscapes as maps of functional diversity in protein sequence space. *Nature Communications*, 14, 2023.