Downscaling metawebs: propagation of uncertainties in species distribution and interaction probability

Gabriel Dansereau [1,2,‡], Ceres Barros [3], Timothée Poisot [1,2]

[1] Université de Montréal; [2] Québec Centre for Biodiversity Sciences; [3] University of British Columbia
[‡] Equal contributions

Correspondance to:
Gabriel Dansereau — gabriel.dansereau@umontreal.ca

1 Introduction

Here, we present a method to downscale a metaweb in space by developing an explicit spatial probabilistic metaweb for Canadian mammals. We present how the spatial structure of the downscaled metaweb varies in space and how the uncertainty of interactions can be made spatially explicit. We further show that the downscaled metaweb can highlight important biodiversity areas and bring novel ecological insights compared to community measures.

2 Methods

Fig. 1 shows a conceptual overview of the methodological steps leading to the downscaled metaweb. The components were grouped as the inputs (spatial or non-spatial), the localized steps (divided into single-species-level, two-species-level, and network-level steps), and the final downscaled and spatialized output. Throughout these steps, we highlight the importance of presenting the uncertainty of both interactions and their distribution in space. We argue that this requires adopting a probabilistic view and incorporating variation between scales.

2.1. Inputs The inputs were divided into two main categories: the spatial and non-spatial ones (*Inputs* box on Fig. 1).

2.1.1 Non-spatial inputs The main building block for the interaction data was the metaweb for Canadian mammals from Strydom *et al.* (2022a), a non-spatial input (represented as nodes and links on Fig. 1). A metaweb contains all the possible interactions between the species found in a given regional species pool (Dunne 2006). The species list for the Canadian metaweb was extracted from the International Union for the Conservation of Nature (IUCN) checklist (Strydom *et al.* 2022a). Briefly, the metaweb was developed using graph embedding and phylogenetic transfer learning based on the metaweb of European mammals, which is itself based on a comprehensive survey of interactions reported in the scientific literature (Maiorano *et al.* 2020). The Canadian metaweb is probabilistic, which has the advantage of taking into account that species do not necessarily interact whenever they co-occur (Blanchet *et al.* 2020). However, the Canadian metaweb is not explicitly spatial: it only gives information on interactions in Canada as a whole and does

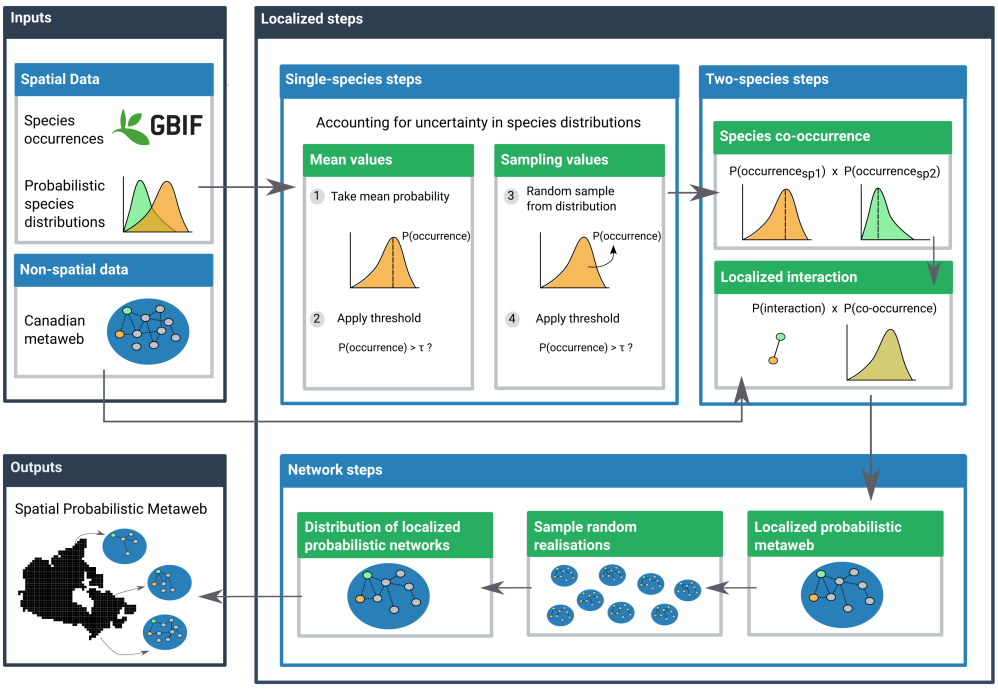

Figure 1 Conceptual figure of the workflow to obtain the spatial probabilistic metaweb (Chapter 1). The workflow has three components: the inputs, the localized steps, and the final spatial output. The inputs are composed of the spatial data (data with information in every cell) and the non-spatial data (constant for all of Canada). The localized steps use these data and are performed separately in every cell, first at a single-species level (using distribution data), then for every species pair (adding interaction data from the metaweb), and finally at the network level by combining the results of all species pairs. The final output coming out of the network-level steps contains a spatialized probabilistic metaweb for every cell across the study extent.

not represent networks at specific locations. Local networks, on the other hand, are realizations from the metaweb resulting from sorting the species and the interactions (Poisot *et al.* 2015). A spatial and localized metaweb is not equivalent to the local networks, as it will have a different structure and a higher connectance (Strydom *et al.* 2022b). Therefore, producing a spatial metaweb requires additional steps to account for species composition and interaction variability in space.

2.1.2 Spatial inputs The spatial data used to develop the spatial component of the metaweb were species occurrences and environmental data. First, we extracted species occurrences from the Global Biodiversity Information Facility (GBIF; www.gbif.org) for the Canadian mammals after reconciling species names between the Canadian metaweb and GBIF using the GBIF Backbone Taxonomy (GBIF Secretariat 2021). Doing so, we removed potential duplicates where species listed in the Canadian metaweb are considered as a single species by GBIF. We collected occurrences for our species list using the GBIF download API on October 21st 2022 (GBIF.org 2022). We restricted our query to occurrences with coordinates between longitudes 175°W to 45°W and latitudes 10°N to 90°N. This was meant to collect training data covering a broader range than our prediction target (Canada only) and include observations in similar environments. Then, since GBIF observations represent presence-only data and most predictive models require absence data, we generated pseudo-absence data using the surface range envelope method available in `SimpleSDMLayers.jl` (Dansereau & Poisot 2021). This method generates pseudo-absences by selecting random non-observed locations within the spatial range delimited by the presence data (Barbet-Massin *et al.* 2012).

We used environmental data and species distribution models (SDMs, Guisan & Thuiller 2005) to predict the distribution of Canadian mammals across the whole country. The environmental data we used were the 19 bioclimatic variables from CHELSA (Karger *et al.* 2017) and the 12 consensus land cover variables from EarthEnv (Tuanmu & Jetz 2014). The CHELSA bioclimatic variables (*bio1-bio19*) represent various measures of temperature and precipitation (e.g., annual averages, monthly maximum or minimum, seasonality) and are available for land areas across the globe. Therefore, they can be used to capture the climatic tolerance of species and model habitat suitability in new locations. We used the most recent version, the CHELSA v2.1 dataset (Karger *et al.* 2021). However, this version also includes bioclimatic data for open water, while we decided here to focus only on land surfaces. We used the previous version, CHELSA v1.2 (Karger *et al.* 2018), which shares a similar grid but does not cover open water, as a mask to clip the v2.1 data to land surfaces only. The EarthEnv land cover variables represent classes such as Evergreen broadleaf trees, Cultivated and managed vegetation, Urban/Built-up, and Open Water. Values range between 0 and 100 and represent the consensus prevalence of each class in percentage within a pixel. We coarsened both the CHELSA and EarthEnv data from their original 30 arc-second resolution to a 2.5 arc-mininute one (around

4.5 km at the Equator) using GDAL/OGR contributors (2021). This represented a compromise to catch both local variations and broad scale patterns while limiting computation costs to a manageable level, as memory requirements on localized interactions rise very quickly.

Our selection criteria for choosing an SDM algorithm was to have a method that generated probabilistic results, including both a probability of occurrence for a species in a specific location and the uncertainty associated with the prediction. These were crucial to obtaining a probabilistic version of the metaweb as they were used to create spatial variations in the localized interaction probabilities (see next section). One promising method for this is Gradient Boosted Trees with a Gaussian maximum likelihood from the `EvoTrees.jl` *Julia* package (https://github.com/Evovest/EvoTrees.jl). This method returns a prediction for every pixel with an average value and a standard deviation, which we used as a measure of uncertainty to build a Normal distribution for the probability of occurrence of a given species at all pixels (represented as probability distributions on Fig. 1). We trained models across the extent chosen for occurrences (longitudes 175°W to 45°W and latitudes 10°N to 90°N), then predicted species distributions only for Canada. We used the 2021 Census Boundary Files from Statistics Canada (Statistics Canada 2022) to set the boundaries for our predictions.

2.2. Localized steps The next part of the method was the localized steps which produce local metawebs in every pixel. This component was divided into single-species, two-species, and network-level steps (*Localized steps* box on Fig. 1).

The single-species steps represented four possible ways to account for uncertainty in the species distributions and bring variation to the spatial metaweb. We explored four different options to select a value from the occurrence distributions obtained in the previous steps (Inputs section): 1) taking the mean from the distribution as the probability of occurrence (option 1 on Fig. 1); 2) converting the mean value to a binary one using a specific threshold per species (option 2); 3) sampling a random value within the Normal distribution (option 3); 4) converting the random value into a binary result (option 4). The threshold (τ on Fig. 1) used was the value that maximized Youden's *J* informedness statistic (Youden 1950), the same metric used by Strydom *et al.* (2022a) at an intermediate step while building the metaweb. The four sampling options were intended to explore how uncertainty and variation in the species distributions can affect the metaweb result and reproduce some of the filterings that create the local network realizations (Poisot *et al.* 2015). We expected thresholding to have a more pronounced effect on network structure as it should reduce the number of links by removing many of the rare interactions (Poisot *et al.* 2016). Meanwhile, we expected random sampling to create spatial heterogeneity compared to the mean probabilities, as including some extreme values should disrupt the potential effects of environmental gradients.

Next, the two-species steps aimed to give the probability of observing a given interaction in a location. For all species pairs, we multiplied the two species' occurrence probability obtained using the sampling options described in the previous paragraph, then multiplied the co-occurrence probability by the interaction probability from the Canadian metaweb. For cases where species in the Canadian metaweb were considered as the same species by the GBIF Backbone Taxonomy (the reconciliation step mentioned earlier), we used the highest interaction probabilities involving the duplicated species.

The network-level steps then created the probabilistic metaweb for the location. We assembled all the local interaction probabilities (from the two-species steps) into a probabilistic network (Poisot *et al.* 2016). We then sampled several random network realizations to represent the potential local realization process (Poisot *et al.* 2015). Finally, this resulted in a distribution of localized networks, which we averaged over the number of simulations to obtain a probabilistic network.

2.3. Outputs The final output of our method was the spatial probabilistic metaweb, which contains a localized probabilistic metaweb in every cell across the student extent (Outputs box on Fig. 1). This gives us an idea of the possible networks in all locations as the metaweb essentially serves to set an upper bound on the potential interactions (Strydom *et al.* 2022b), but with the added benefit of accounting for co-occurrence probabilities in this case. From there, we can create maps of network properties (e.g. number of links, connectance) measured on the local realizations, display their spatial distribution, and compute some community-level measures such as species richness. We can also calculate the uncertainty associated with the network and community measurements and contrast their spatial distribution (see Supplementary Material).

2.3.1 Ecoregions Since both species composition and network summary values display a high spatial variation and complex patterns, we simplified the representation of their distribution by grouping sites by ecoregion, as species and interaction composition have been shown to differ between ecoregions across large

spatial scales (Martins *et al.* 2022). To do so, we used the global map of ecoregions from (Dinerstein *et al.* 2017; also used by Martins *et al.* 2022), rasterized it, and clipped it to Canada, which selected 44 different ecoregions. For every measure we report (e.g. species richness, number of links), we first calculated the measure for every site separately, then we extracted the median value for each ecoregion. We also measured the within-ecoregion variation by measuring the 89% interquantile range of the values in each ecoregion (threshold chosen to avoid confusion with conventional significance tests, inspired by McElreath 2020).

2.3.2 Ecological uniqueness We compared the compositional uniqueness of the networks and the communities to verify if they indicated different exceptional areas. We measured uniqueness using the local contributions to beta diversity (LCBD, Legendre & De Cáceres 2013), which identify sites with exceptional composition by quantifying how much one site contributes to the total variance in the community composition. While many studies used LCBD values to evaluate uniqueness on local scales or few study sites (for example, da Silva & Hernández 2014; Heino & Grönroos 2017), recent studies used the measure on predicted species compositions over broad spatial extents and a large number of sites (Vasconcelos *et al.* 2018; Dansereau *et al.* 2022). LCBD values can also be used to measure uniqueness for networks by computing the values over the adjacency matrix, which has been shown to capture more unique sites and uniqueness variability than through species composition (Poisot *et al.* 2017). Here, we measured and compared the uniqueness of our localized community and network predictions. We were especially interested in seeing if the sites identified as unique were the same based on the species and the interactions or if this method allowed identifying areas unique for one element (interactions, for instance) but not the other. Sites with such mismatches should warrant more investigation to understand the reasons for this difference.

2.4. Software We used *Julia* v1.9.0 (Bezanson *et al.* 2017) to implement all our analyses. We used packages `GBIF.jl` (Dansereau & Poisot 2021) to reconcile species names using the GBIF Backbone Taxonomy, `SpeciesDistributionToolkit.jl` to handle raster layers and species occurrences, `EcologicalNetworks.jl` (Poisot *et al.* 2019) to analyse network and metaweb structure, and `Makie.jl` (Danisch & Krumbiegel 2021) to produce figures. Our data sources (CHELSA, EarthEnv, Ecoregions) were all unprojected and we did not use a projection in our analyses, but we displayed the results using a Lambert conformal conic projection more appropriate for Canada using `GeoMakie.jl`. All the code used to implement our analyses is available on GitHub (https://github.com/PoisotLab/SpatialProbabilisticMetaweb) and includes instructions on how to run a smaller example at a coarse resolution. Note that running our analyses at full scale is resource and memory intensive and required the use of compute clusters provided by Calcul Québec and the Digital Research Alliance of Canada.

3 Results

Our method allowed us to display the spatial distribution of ecoregion-level measures (Fig. 2), either for community measures (e.g. expected species richness) or for network measures (e.g. expected number of links). Importantly, both community and network-level measures presented are not predictions of the measure itself but were instead computed over localized predictions of the communities and networks, then summarized for the ecoregions. Expected ecoregion richness (Fig. 2A), which is the median of the expected species richness of the sites within the ecoregion, and expected number of links (Fig. 2B) displayed similar distributions with a latitudinal gradient and higher values in the south. However, within-ecoregion variability was distributed differently, as some ecoregions along the coasts displayed higher interquantile ranges while ecoregions around the southern border displayed narrower ones (Fig. 2C-D). All results shown are based on the first sampling strategy (option 1) mentioned in the Localized steps section, where species occurrence probabilities were taken as the mean value of the distribution (results for other sampling strategies are discussed in Supplementary Material).

Direct comparison of the spatial distributions of species richness and expected number of links showed some areas with mismatches, both regarding the median estimates and regarding the within-ecoregion variability Fig. **??**. Median values for the ecoregions showed a similar bivariate distribution with ecoregions in the south mostly displaying high species richness and a high number of links (Fig. **??**A). The northernmost ecoregions (Canadian High Artic Tundra and Davis Highlands Tundra) displayed higher richness (based on the quantile rank) compared to the number or links. Inversely, ecoregions further south (Canadian Low Artic Tundra, Northern Canadian Shield Taiga, Southern Hudson Bay Taiga) ranked higher for the number of links than for species richness. On the other hand, within-ecoregion variability showed different bivariate relationships and

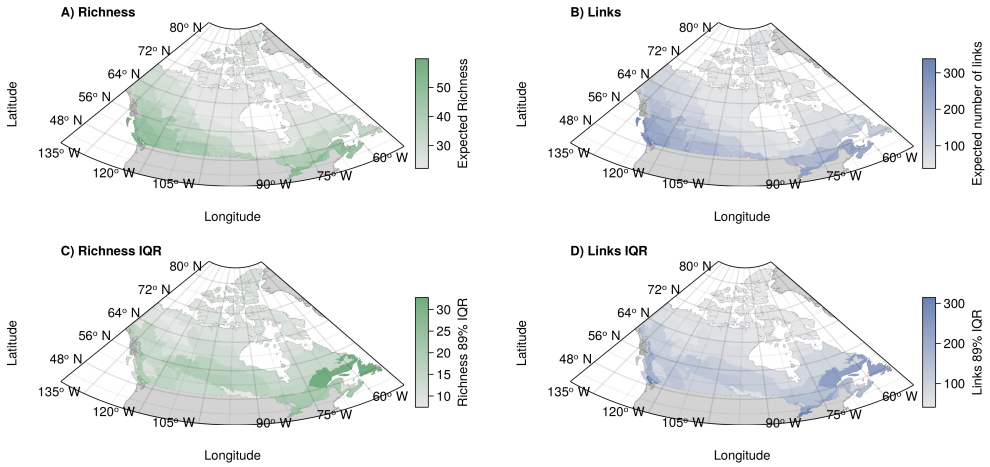

Figure 2 (A-B) Example of a community measure (A, expected species richness) and a network one (B, expected number of links). Both measures are assembled from the predicted probabilistic communities and networks, respectively. Values are first measured separately for all sites, then the median value is taken to represent the ecoregion-level value. (C-B) Representation of the 89% interquantile range of values within the ecoregion for expected richness (C) and expected number of links (D).

a less constant latitudinal gradient. This indicates that richness and link do not vary hand in hand (i.e. their variability is not closely connected) although they may show similar distributions for median values.

Our results also indicate a mismatch between the uniqueness of communities and networks (Fig. 4). Uniqueness was higher mostly in the north and along the south border for communities, but only in the north for networks (Fig. 4A-B). Consequently, ecoregions with both unique community composition and unique network composition were mostly in the north (Fig. 4C). Meanwhile, some areas were unique for one element but not the other. For instance, the New England-Acadian forests ecoregion (south-east, near 70°W and 48°N) had a highly unique species composition (Fig. 4C) but a more common network composition. Opposite areas with unique network compositions only were higher north between latitudes 52°N and 70°N (Eastern Canadian Shield Taiga, Northern Canadian Shield Taiga, Canadian Low Artic Tundra). When comparing the values, network uniqueness values for ecoregions spanned a narrower range between the 44 ecoregions than species LCBD values (Fig. 4D, left). Within-ecoregion variation was also lower for network values with generally lower 89% interquantile ranges among the site-level LCBD values (Fig. 4D, right). Moreover, mismatched sites (unique for only one element) formed two distinct groups when evaluating the relationship between species richness and the number of links (see Supplementary Material). The areas only unique for their species composition had both a high richness and number of links. On the other hand, the sites only unique for their networks had both lower richness and a lower number of links, although they were not the sites with the lowest values for both.

Discussion

Barbet-Massin, M., Jiguet, F., Albert, C.H. & Thuiller, W. (2012). Selecting pseudo-absences for species distribution models: How, where and how many? *Methods in Ecology and Evolution*, 3, 327–338.

Bezanson, J., Edelman, A., Karpinski, S. & Shah, V.B. (2017). Julia: A fresh approach to numerical computing. *SIAM Review*, 59, 65–98.

Blanchet, F.G., Cazelles, K. & Gravel, D. (2020). Co-occurrence is not evidence of ecological interactions. *Ecology Letters*, 23, 1050–1063.

da Silva, P.G. & Hernández, M.I.M. (2014). Local and regional effects on community structure of dung beetles in a mainland-island scenario. *PLOS ONE*, 9, e111883.

Danisch, S. & Krumbiegel, J. (2021). Makie.jl: Flexible high-performance data visualization for Julia. *Journal of Open Source Software*, 6, 3349.

Dansereau, G., Legendre, P. & Poisot, T. (2022). Evaluating ecological uniqueness over broad spatial extents using species distribution modelling. *Oikos*, 2022, e09063.

Dansereau, G. & Poisot, T. (2021). SimpleSDMLayers.jl and GBIF.jl: A framework for species distribution modeling in Julia. *Journal of Open Source Software*, 6, 2872.

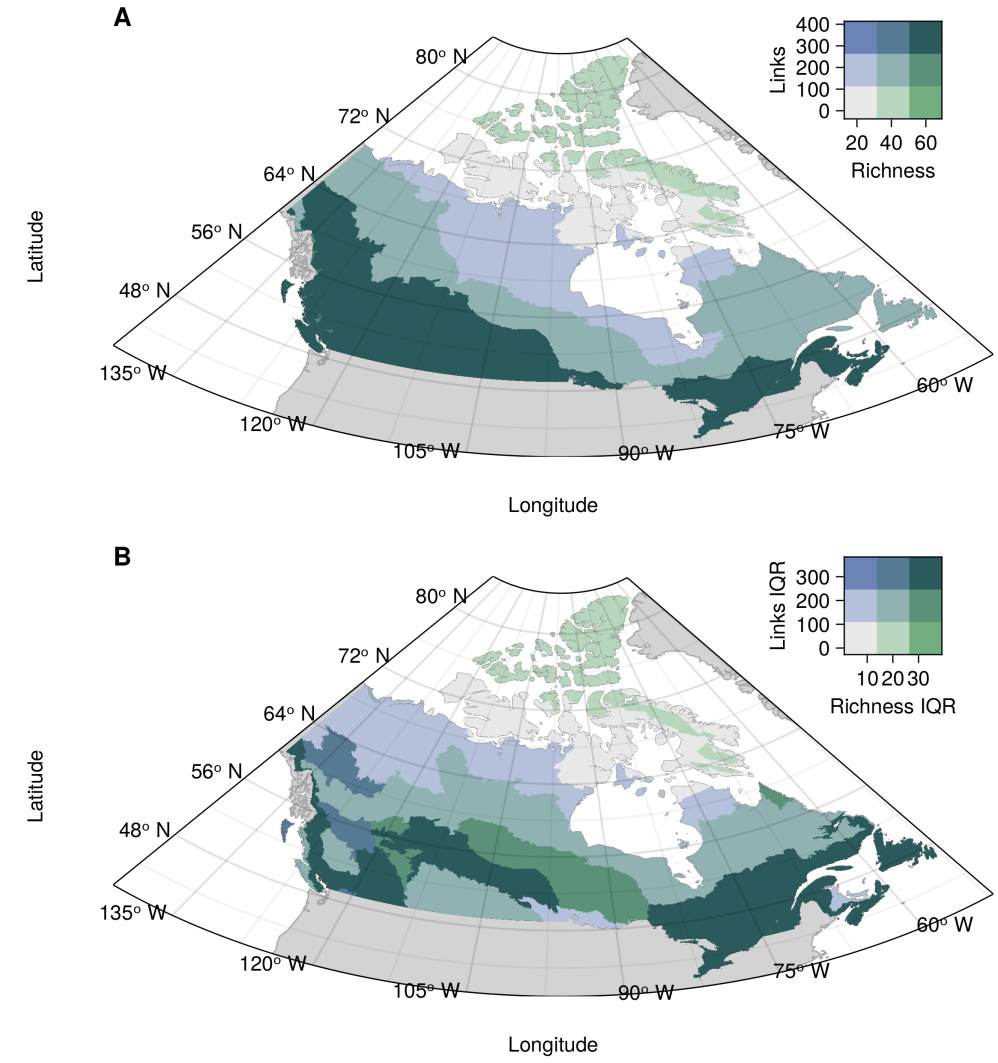

Figure 3 Bivariate relationship between community and network measures for the median ecoregion value (A) and the within-ecoregion 89% interquantile range (B). Values are grouped into three quantiles separately for each variable. The colour combinations represent the nine possible combinations of quantiles. Species richness (horizontal axis) goes left to right from low (light grey, bottom left) to high (green, bottom right). The number of links goes bottom-up from low (light grey, bottom left) to high (blue, top left).

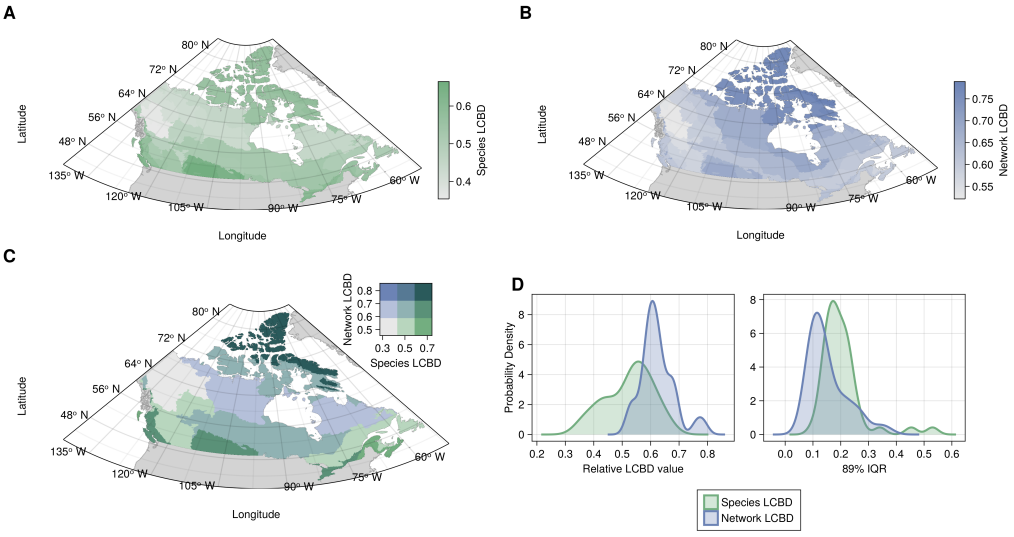

Figure 4 (A-B) Representation of the ecoregion uniqueness values based on species composition (a) and network composition (b). LCBD values were first computed across all sites and scaled relative to the maximum value observed. The ecoregion LCBD value is the median value for the sites in the ecoregion. (C) Bivariate representation of species and network composition LCBD. Values are grouped into three quantiles separately for each variable. The colour combinations represent the nine possible combinations of quantiles. The species uniqueness (horizontal axis) goes left to right from low uniqueness (light grey, bottom left) to high uniqueness (green, bottom right). The network composition uniqueness goes bottom-up from low uniqueness (light grey, bottom left) to high uniqueness (blue, top left). (D) Probability densities for the ecoregion LCBD values for species and network LCBD (left), highlighting the variability of the LCBD between ecoregions, and the 89% interquartile range of the values within each ecoregion (right), highlighting the variability within the ecoregions.

Dinerstein, E., Olson, D., Joshi, A., Vynne, C., Burgess, N.D., Wikramanayake, E., *et al.* (2017). An Ecoregion-Based Approach to Protecting Half the Terrestrial Realm. *BioScience*, 67, 534–545.

Dunne, J. (2006). The network structure of food webs. In: *Ecological Networks: Linking Structure to Dynamics in Food Webs*. pp. 27–86.

GBIF Secretariat. (2021). GBIF Backbone Taxonomy.

GBIF.org. (2022). GBIF occurrence download.

GDAL/OGR contributors. (2021). *GDAL/OGR geospatial data abstraction software library*. Manual. Open Source Geospatial Foundation.

Guisan, A. & Thuiller, W. (2005). Predicting species distribution: Offering more than simple habitat models. *Ecology Letters*, 8, 993–1009.

Heino, J. & Grönroos, M. (2017). Exploring species and site contributions to beta diversity in stream insect assemblages. *Oecologia*, 183, 151–160.

Karger, D.N., Conrad, O., Böhner, J., Kawohl, T., Kreft, H., Soria-Auza, R.W., *et al.* (2017). Climatologies at high resolution for the earth's land surface areas. *Scientific Data*, 4, 170122.

Karger, D.N., Conrad, O., Böhner, J., Kawohl, T., Kreft, H., Soria-Auza, R.W., *et al.* (2018). Data from: Climatologies at high resolution for the earth's land surface areas.

Karger, D.N., Conrad, O., Böhner, J., Kawohl, T., Kreft, H., Soria-Auza, R.W., *et al.* (2021). Climatologies at high resolution for the earth's land surface areas.

Legendre, P. & De Cáceres, M. (2013). Beta diversity as the variance of community data: Dissimilarity coefficients and partitioning. *Ecology Letters*, 16, 951–963.

Maiorano, L., Montemaggiori, A., Ficetola, G.F., O'Connor, L. & Thuiller, W. (2020). TETRA-EU 1.0: A species-level trophic metaweb of European tetrapods. *Global Ecology and Biogeography*, 29, 1452–1457.

Martins, L.P., Stouffer, D.B., Blendinger, P.G., Böhning-Gaese, K., Buitrón-Jurado, G., Correia, M., *et al.* (2022). Global and regional ecological boundaries explain abrupt spatial discontinuities in avian frugivory interactions. *Nature Communications*, 13, 6943.

McElreath, R. (2020). *Statistical rethinking: A bayesian course with examples in R and Stan*. Second. Chapman and Hall/CRC, New York.

Poisot, T., Bélisle, Z., Hoebeke, L., Stock, M. & Szefer, P. (2019). EcologicalNetworks.jl: Analysing ecological networks of species interactions. *Ecography*, 42, 1850–1861.

Poisot, T., Cirtwill, A.R., Cazelles, K., Gravel, D., Fortin, M.-J. & Stouffer, D.B. (2016). The structure of probabilistic networks. *Methods in Ecology and Evolution*, 7, 303–312.

Poisot, T., Guéveneux-Julien, C., Fortin, M.-J., Gravel, D. & Legendre, P. (2017). Hosts, parasites and their interactions respond to different climatic variables. *Global Ecology and Biogeography*, 26, 942–951.

Poisot, T., Stouffer, D.B. & Gravel, D. (2015). Beyond species: Why ecological interaction networks vary through space and time. *Oikos*, 124, 243–251.

Statistics Canada. (2022). *Boundary files, reference guide second edition, Census year 2021*. Second edition. Statistics Canada = Statistique Canada, Ottawa.

Strydom, T., Bouskila, S., Banville, F., Barros, C., Caron, D., Farrell, M.J., *et al.* (2022a). Food web reconstruction through phylogenetic transfer of low-rank network representation. *Methods in Ecology and Evolution*, n/a.

Strydom, T., Bouskila, S., Banville, F., Barros, C., Caron, D., Farrell, M.J., *et al.* (2022b). Predicting metawebs: Transfer of graph embeddings can help alleviate spatial data deficiencies.

Tuanmu, M.-N. & Jetz, W. (2014). A global 1-km consensus land-cover product for biodiversity and ecosystem modelling. *Global Ecology and Biogeography*, 23, 1031–1045.

Vasconcelos, T.S., Nascimento, B.T.M. do & Prado, V.H.M. (2018). Expected impacts of climate change threaten the anuran diversity in the Brazilian hotspots. *Ecology and Evolution*, 8, 7894–7906.

Youden, W.J. (1950). Index for rating diagnostic tests. *Cancer*, 3, 32–35.