# Proof of concept

As a proof-of-concept of multivariate constraint, we developed a simple joint analysis of vegetation aboveground biomass and population density of moose (*Alces alces*) in North America. This analysis uses information from a combination of models and observations and demonstrates how a model that jointly predicts moose density and vegetation biomass allows us to use information on one variable to indirectly constrain others.

## Data descriptions and processing

First, we obtained aboveground vegetation biomass estimates from the North American Carbon Program (NACP) Multi-scale Synthesis and Terrestrial Model Intercomparison Project (MstMIP; Huntzinger et al. 2018). The complete dataset comprises global gridded 0.5° x 0.5° simulations from 15 different Terrestrial Biosphere Models at monthly and yearly time steps for the period 1900-2010. For each model and each pixel within North America, we calculated the mean biomass estimate for the period 1990-2010 for each pixel. We then generated a normal distribution of biomass for each pixel by taking the sample mean and standard deviation across the 15 models. A map of the resulting mean estimates is shown in the top panel of Figure 1.

Second, we obtained estimates of vegetation biomass and moose population density from existing global gridded ensemble simulations of the Madingley model (Harfoot et al. 2014). The ensemble was comprised of six different simulations with varying parameters and initial conditions. To get vegetation biomass, we first extracted the predicted “autotroph biomass density” from each ensemble member for each pixel in North America. In Madingley, this quantity corresponds to the total foliar biomass; to convert to total biomass, we multiplied these values by 10, based on allometric relationships between foliar and total biomass in Bond-Lamberty et al. (2002). A map of the resulting mean estimates of vegetation biomass is shown in the middle panel of Figure 1. To get moose density, we first extracted the predicted adult population of each cohort of terrestrial herbivores whose body mass was greater than 400 kg for each pixel. (Madingley does not differentiate species beyond diet and body mass. Our criteria effectively exclude the other two largest wild herbivores in North America: White-tailed deer [*Odocoileus virginianus*]—whose adult mass is typically less than 200 kg—and most subspecies of elk [*Cervus canadensis*]—whose adult mass is typically less that 350 kg. Adult mass of Roosevelt elk [*C. canadensis roosevelti*] can reach as high as 600 kg, but their geographic extent is relatively limited compared to moose). We then calculated the population density within each pixel by taking the sum of the populations of all cohorts meeting the criteria above and dividing by the area of each pixel. A map of the resulting mean estimates of moose density is shown in the middle panel of Figure 2.

Finally, we obtained a separate gridded estimate of moose density from XXX. We first converted this shapefile of moose density to a raster matching the grid of the above using the `rasterize` function in the R package `raster` (Hijmans 2020). This dataset did not come with uncertainty estimates, so we assumed a uniform relative standard deviation of 10% of the stated value for every pixel.

## Analysis

For each of the four datasets above—vegetation biomass from MstMIP and Madingley, and moose density from Madingley and XXX—we have rasters of means and standard deviations from each pixel. The mean rasters are shown in the top two panels of Figures 1 and 2. For an initial evaluation of agreement between these estimates, we calculated the probability of overlap of the corresponding distributions as follows: For each pixel and each variable, we calculated the minimum of the corresponding probability density function at every point and numerically integrated the resulting values over all regions of non-trivial probability density:

The resulting gridded estimates of overlap for vegetation biomass and moose density are shown in the bottom panels of Figures 1 and 2, respectively.

We performed two different kinds of constraint on the output variables: A univariate constraint where moose density and biomass were each constrained independently of one another, and a multivariate constraint whereby they were constrained jointly. In the univariate constraint, because we assume both variables are normally distributed, we can analytically calculate the best estimate of each variable (*Xbest*) from the distributions of its estimates:

The results of the independent univariate constraints on each variable are shown in Figure 3.

The multivariate distribution of Madingley ensemble members was not well approximated by a multivariate normal distribution. Therefore, we performed this constraint numerically, by weighting each Madingley ensemble member according to its agreement with MstMIP vegetation biomass and observed moose density. Let and be the estimates of moose density and biomass, respectively, by Madingley ensemble member . Also, let and be the mean and standard deviation of the distribution of MstMIP biomass estimates, and be the analogous quantities for observed moose biomass, and be the number of Madingley ensemble members. The probability of Madingley ensemble member is therefore given by:

We define the weight of each ensemble member as the normalized probability:

Based on these quantities, we can calculate the jointly constrained distribution of biomass as:

Analogous equations can be used for moose density.

For the joint constraint to produce meaningful and useful results, the distributions of the underlying variables need to have non-trivial overlap for both variables of interest. While this was often true for vegetation biomass (Figure 1), overlap in moose density estimates was poor in all but a few specific locations (Figure 2), resulting in poor joint constraint results in most places. Therefore, we show the joint constraint analysis described above for one specific pixel.

We recognize that the above procedure includes many simplistic assumptions, both scientifically and statistically, and therefore that the results of this analysis are not meant to be robust or reliable. Rather, our objective was to show a simple example of how this might work. In real settings, more careful consideration should be given to the choice of distributions (or, non-parametric techniques such as bootstrapping can be used) and to the processing steps used to align different variables.

|  |
| --- |
| A close up of a map  Description automatically generated  **Figure 1:** Mean estimates of vegetation biomass from the MstMIP multi-model ensemble (top) and Madingley model ensemble (middle), and the fraction of distribution overlap of these estimates and their uncertainties for each pixel (bottom; see methods for details). |

|  |
| --- |
| **A close up of a map  Description automatically generated**  **Figure 2:** Same as Figure 1, but for moose density from observations (top) and the Madingley model (middle). Note that, unlike Figure 1, the top and middle panels have different scales. The pixel used for the joint constraint analysis in Figures 3 and 4 is indicated by the symbol in the bottom panel. |

|  |
| --- |
| **A screenshot of a map  Description automatically generated**  **Figure 3:** Univariate constraint analysis, as described in methods. Black line is the distribution of MstMIP ensemble estimates of vegetation biomass (left) and observed moose density (right). Red line is the distribution of both variables from Madingley, with labeled red dashes showing individual ensemble members. Green line shows the best-estimate distribution from combining the two estimates, as described in the methods. |

|  |
| --- |
| **A close up of a map  Description automatically generated**  **Figure 4:** Multivariate constraint analysis. Labeled black dots are Madingley ensemble joint estimates of moose density and vegetation biomass. Shading indicates the joint probability of moose density and vegetation biomass based on the MstMIP/observed distributions (black lines in the marginal density plots). Blue line in marginal density plot represents the marginal distribution of each variable after the joint constraint (see methods) has been applied. Green dashed line in the marginal plots is the univariate constraint estimate, as in Figure 3. |

# References

Bond-Lamberty, B, C Wang, and S T Gower. “Aboveground and Belowground Biomass and Sapwood Area Allometric Equations for Six Boreal Tree Species of Northern Manitoba.” Canadian Journal of Forest Research 32, no. 8 (August 2002): 1441–50. <https://doi.org/10.1139/x02-063>.

Harfoot, Michael B. J., Tim Newbold, Derek P. Tittensor, Stephen Emmott, Jon Hutton, Vassily Lyutsarev, Matthew J. Smith, Jörn P. W. Scharlemann, and Drew W. Purves. “Emergent Global Patterns of Ecosystem Structure and Function from a Mechanistic General Ecosystem Model.” Edited by Michel Loreau. PLoS Biology 12, no. 4 (April 22, 2014): e1001841. <https://doi.org/10.1371/journal.pbio.1001841>.

Hijmans, Robert J. (2020). raster: Geographic Data Analysis and Modeling. R package version 3.0-12. https://CRAN.R-project.org/package=raster

Huntzinger, D.N., C.R. Schwalm, Y. Wei, R.B. Cook, A.M. Michalak, K. Schaefer, A.R. Jacobson, M.A. Arain, P. Ciais, J.B. Fisher, D.J. Hayes, M. Huang, S. Huang, A. Ito, A.K. Jain, H. Lei, C. Lu, F. Maignan, J. Mao, N.C. Parazoo, C. Peng, S. Peng, B. Poulter, D.M. Ricciuto, H. Tian, X. Shi, W. Wang, N. Zeng, F. Zhao, Q. Zhu, J. Yang, and B. Tao. 2018. NACP MsTMIP: Global 0.5-degree Model Outputs in Standard Format, Version 1.0. ORNL DAAC, Oak Ridge, Tennessee, USA. <https://doi.org/10.3334/ORNLDAAC/1225>