Supporting Information:

Do people manage climate risk through long-distance relationships?

Anne Pisor & James Holland Jones

Version: June 30, 2020

Data, metadata, and code available at XXX

CONTENTS

1. COLLABORATING COMMUNITIES

The two collaborating communities of horticulturalists are introduced in the main text (Section 2.1). Here, we confine our discussion to additional relevant details that supplement the main text.

The first community, a multicultural (*intercultural*) community, was founded in the early 1970s by families who had come to the area to work in the quinine or who moved there as part of a voluntary government relocation program. The plurality of community members is first- or second-generation emigrants from the Bolivian highlands. After first relying solely on river transit, pack mules, or travel by foot to reach the local market town, the community gained road access in 1975 (Llojlla Roca, 2011), facilitating direct sales of cash crops to middlemen, short-term migration, and further emigration of families interested in horticulture or logging. This Intercultural community is located one hour's drive away from the market town and seven hours away from La Paz, the de facto national capital and largest nearby city.

The second community was founded by Mosetén families, a local indigenous group. As the Mosetén community at a local mission grew larger in the latter part of the 20th century, a shortage of land forced families to establish horticultural plots further and further away from the mission community; in the late 1990s, they founded a new community to bring services (including a school and health care) closer to their homes. In 2000, a stone-paved road to the community was completed, facilitating crop sales, short-term migration, and emigration (Pisor & Gurven, 2018). This Mosetén community is one and a half hours from the local town and seven and a half hours from La Paz.

2. PRECIPITATION DATA

Precipitation accumulation data were downloaded from the NCAR Computational and Information Systems Laboratory (*Standardardized Precipitation Index (SPI) for Global Land Surface (1949-2012)*, 2013). These data are discussed in the main text; here, we confine our discussion to additional relevant details that supplement the main text.

SPI data were available for a range of different windows, from three months to 48 months in length. Each of these window lengths is useful for different purposes; for example, three month windows provide more insight into how precipitation affects agriculture, while 48 month windows provide more insight into how precipitation affects reservoirs or water tables (Svoboda et al., 2012). Given our interest in the impact of climate variability on human livelihoods, we focus on three-month windows: for the two communities in Bolivia, even one month of high precipitation can cause the failure of crucial cash crops, including papaya, cacao, and yuca.

2.1 Additional limitations of the precipitation data

Daily precipitation accumulation tends to follow a gamma distribution (Martinez-Villalobos & Neelin, 2019). However, common practice in meteorological science is to average daily precipitation accumulation by month or by longer periods, up to 48 months in length, and then normalize these gamma-distributed data into a standard normal distribution, called the Standard Precipitation Index (SPI) (Orlowsky & Seneviratne, 2013). This approximation is appropriate if the shape parameter is large enough, which it usually is for precipitation data. In particularly arid environments, data can be too zero-inflated to be approximated by a gamma distribution (Mishra & Singh, 2010). However, only one department in our data set -- Tarapacá, in Chile -- has many zero values for precipitation, and only two participants are coded as having lived in Tarapacá. As such, the potential effects of an inappropriate conversion of gamma to standardized normal in the case of Tarapacá is unlikely to affect our results.

The NCAR dataset did not include data from before 1949. We attempted to remedy this problem with raw data of daily weather station precipitation data from NOAA, however there are few data available for Bolivian stations before the 1980s. We are thus unable to account for participants' exposure to precipitation if they were born before 1949; this affects a total of three participants. The NCAR dataset also did not include data from after 2012, affecting all participants in the sample. When we scale by months of exposure for testing H3, we thus scale only by the years or months of the participant's life for which precipitation data were available.

There is within-department variation in rainfall that is masked when the SPI for a particular month is averaged at the department level. For example, the La Paz Department includes rainforest, cloud forest, and Andean plateaus, all with a very different precipitation profile. However, because AP asked only about department of residence, we do not have finer-grained data with which to assess participants' exposure to precipitation.

3. STATISTICAL MODELS

We pre-registered the analyses reported here on the Open Science Framework: https://osf.io/5pdn3/. AP pre-registered data collection too (https://osf.io/utwuf), although the goals of data collection differed from the hypotheses tested in this paper.

3.1 Sequential models

Under sequential models, a participant cannot reach the next "level" of long-distance relationships without having reached the previous level first: one cannot have three long-distance relationships without first having had two (P.-C. Bürkner & Vuorre, 2019). We use continuous parameterization with our sequential models because we are interested in whether experience of drought or excess precipitation predicts attaining one long-distance relationship instead of zero, or two instead of one – that is, continuing past each level to the next (P.-C. Bürkner & Vuorre, 2019).

3.2 Weakly informative priors

We utilize very weakly informative priors for the models analyzed in this paper. Non-informative priors – commonly, flat priors with high lower and upper bounds – are usually an inappropriate choice for estimating parameter values: given that the null hypothesis is that predictors should not have any relationship with the outcome variable, the sampling process should focus on values closer to zero. To estimate the predictors of interest and third variables, we use the normal distribution with a mean of zero and a standard deviation of 10 as a prior. To estimate the random effect for community – reported as a standard deviation, which must be positive by definition – we use a half-Cauchy distribution with a location parameter of 0 and shape parameter of 2. For discussion of these choices of priors, see McElreath (2020).

3.3 Equidistant thresholds and unequal variance by community

We begin the modeling process by using flexible spacing between thresholds for each level of long-distance relationships. We do this because it may be more difficult for a participant with zero long-distance connections to build one compared to how difficult it is for a participant with two long-distance connections to build a third, meaning that more drought exposure may be required to cross the zero-to-one threshold compared to the two-to-three threshold. However, because assuming equidistance between thresholds involves estimating fewer parameters and thus increasing model power, we assessed whether the assumption of equidistance is appropriate for these models.

We found the assumption of equidistant thresholds (as opposed to flexible thresholds) warranted for all sequential model fits. We compared model fits with flexible thresholds – that is, where the unique spacing between each pair of thresholds is estimated as part of the model fitting process – and models fits with equidistant thresholds – where the *identical* spacing between each pair of thresholds is estimated as part of the model fitting process – using leave-one-out cross-validation (Bürkner & Vuorre, 2019). By leaving out a single observation, refitting, and attempting to predict the value of the outcome variable for the omitted observation, and iterating that process, we can compare the ability of the flexible threshold and equidistant threshold models to predict omitted observations. This comparison is done using the leave-one-out information criterion (LOOIC). LOOIC has advantages over other comparison techniques for Bayesian models, such as the Akaike Information Criterion (AIC), as it incorporates model priors and does not assume that the outcome variable is normally distributed; however, it is less sensitive to the choice of model priors than is the widely applicable information criterion (WAIC), making it a robust choice for our sequential models with weakly informative priors (Vehtari et al., 2017). As is standard practice for AIC, WAIC, and LOOIC, we compared the LOOIC of models with flexible and equidistant thresholds, selecting the model with the lowest LOOIC value.

In ordinal models, it is possible that there is unequal variance in the outcome by group, which can lead to inaccurate model estimates (Bürkner & Vuorre, 2019). Though the inclusion of a random effect for community estimates the amount of variation in the *observed* outcome due to community, it does not account for unequal variance in the latent variable that generated the different levels of friendship. Though this is more commonly a concern for cumulative models than for sequential models (Bürkner & Vuorre, 2019), we conducted exploratory analyses to examine whether unequal variance between communities was a concern. The community-specific intercepts returned by these analyses had CIs that included an OR of 1, suggesting no unequal variance between communities. Further, these models generated unrealistic model estimates with large credible intervals, indicating substantial uncertainty in the model fitting process. As such, we do not report these models in the main text or Supporting Information; however, the code for running them can be found in our GitHub repository.

3.4 Outliers removed during the model fitting process

Exploratory data analysis indicated that outliers in the predictor variables could influence model fit for P1-P3. During the process of assessing different model fits, our first step was to fit a model in which we explicitly estimated the effect of the predictor of interest (e.g., average length of drought intervals or excess precipitation intervals) on the latent outcome; the latent outcome is called "discrimination," or "disc" (Bürkner & Vuorre, 2019). We then examined plots of the predicted values for the latent outcome, disc, to look for the potential influence of outliers; see our GitHub repository for the relevant code.

Outliers appeared to influence fit for the P1 drought model (n=2), as evidenced by both the plot of predicted values and, in the model fit including these participants, reduced estimated sample sizes for some variables and increased uncertainty in parameter estimates compared to a model fit excluding these participants. The P1 drought model reported in the main text (Table 1) excludes these individuals. In our checks of the other four main models, we found no evidence that outliers influenced model fits. See our GitHub repository for reproducible code.

3.5 Procedures to reduce participant identifiability

We took two major precautions to reduce the identifiability of our participants. First, we converted all participant identification numbers (PIDs) to randomized identification numbers (RIDs) so that even the researchers cannot recognize these participants in publicly accessible data. Further, note that these RIDs are different across papers published using this data set; merging data sets from across publications by RID is not possible. Second, after completing the process of model fitting, we binned participant birth years into five-year bins (e.g., 1971 became 1970; 1999 became 1995) to further reduce participant identifiability. The binning of birth years did not alter model results.

3.6 What are odds ratios and credible intervals?

All models discussed in this paper use a logit link function ($\log \frac{p}{1-p}$) to translate the linear combination of predictors, which can have values that range from -/+ infinity, to the ranges appropriate for an ordinal (e.g., 0-3) or binary (i.e., 0, 1) outcome; that means that our models return parameter estimates on the logit scale (also known as log odds). In order to make parameter estimates interpretable, we exponentiate these estimates and report an odds ratio (OR; $\frac{p}{1-p}$) for each parameter. The OR indicates the increased or decreased odds of having one additional long-distance friend (ordinal models) or of having a same-community relationship (binomial models) given a participant's level for a given variable. For example, if the parameter estimate for the average length of a drought interval was 0.5, that would indicate that for every additional month of length of a participant's average drought interval, they would be 0.5 times less likely to have a long-distance friend. Given we are using a Bayesian approach, we report credible intervals (CIs) rather than confidence intervals. CIs indicate the probability that the true model estimate falls within that range. We use 90% CIs here, which indicate that there is a 90% probability that the true estimate falls between the lower bound and the upper bound of the interval. To translate from the log odds (logit) to odds scale, we likewise exponentiate the lower and upper bounds of the CIs; note that because exponentiating converts numbers from a scale of magnitude to a linear scale, CIs are no longer centered around the parameter estimates after exponentiation.

Tables S1a-d. Descriptive statistics

(a) Ordinal variables that include zero in their range

Variable	Zero	One	Two	Three
Long-Distance*	30	35	27	27
Reciprocal Long-Distance*	58	40	21	
Same-Community*	32	87		
Non-Con. Kin Same-Community*	103	16		
Smartphone	46	73		
Vehicle	84	35		

*All refer to relationships (e.g., “long-distance relationships”). Non-con. is non-consanguineal.

(b) Ordinal variables that include one in their range

Variable	One	Two	Three	Four
Extraversion: Stranger*	60	16	42	
Extraversion: Conversation**	22	6	26	64

*This scale was coded as: “Never” = 1, “Sometimes” = 2, “Always” = 3.

**This scale was coded as: “Never” = 1, “Rarely” = 2, “Sometimes” = 3, “Always” = 4. For more information, see metadata available on GitHub.

(c) Nominal variables

Variable	One	Two
Community*	52	67
Sex**	67	52

*Community One is the Intercultural community, Community Two is the Mosetén community.

**Sex One is female, Sex Two is male.

(d) Continuous variables

Variable	Mean	SD
Number Dry Months	47.08	13.87
Number Wet Months	49.91	16.58
Mean Len. Drought	6.39	0.46
Mean Len. Excess P.*	5.80	0.82
Mean Len. No D. or E.P.*	6.75	0.93
Birth Year	1973.92	12.04
Depts. & Countries Visited	1.33	1.36

*P. = precipitation. D. = drought. E.P. = excess precipitation

Figure S1. Parameter estimates and 90% credible intervals for a robustness check of the five main models: treating *reciprocity-based long-distance relationships* as the outcome.

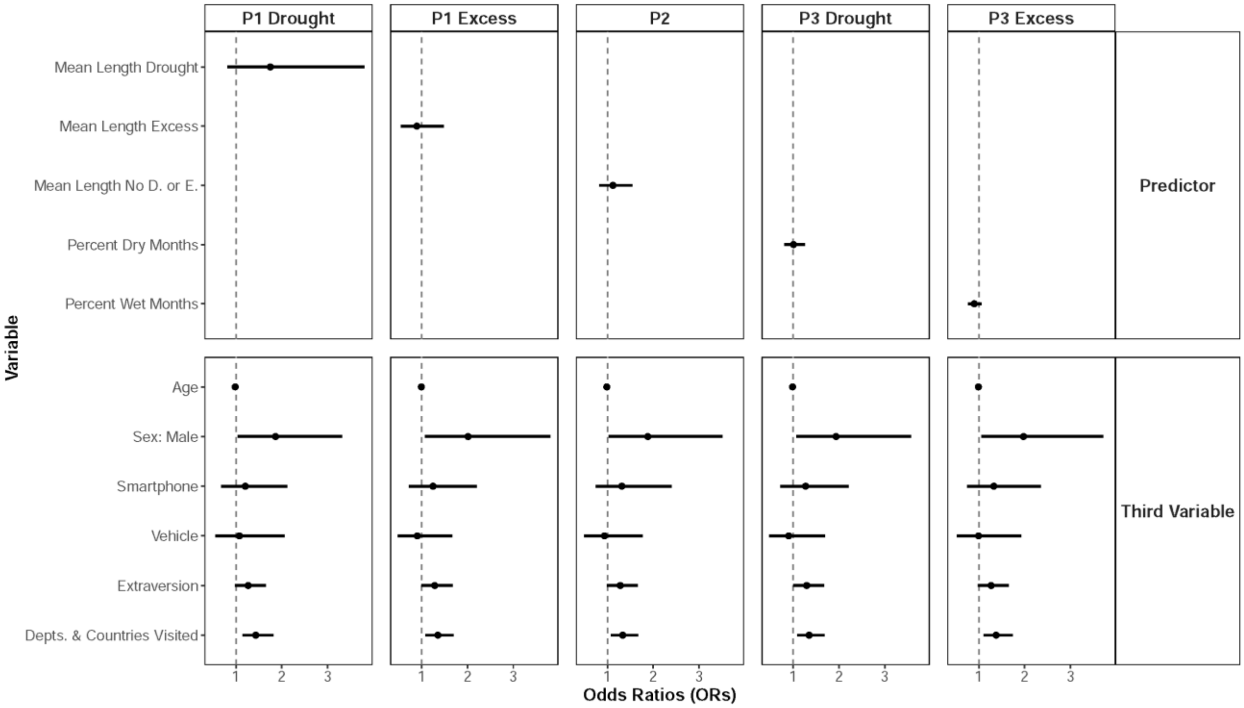


Figure S2. Parameter estimates and 90% credible intervals for an alternative outcome for the five main models: treating *same-community relationships* as the outcome.

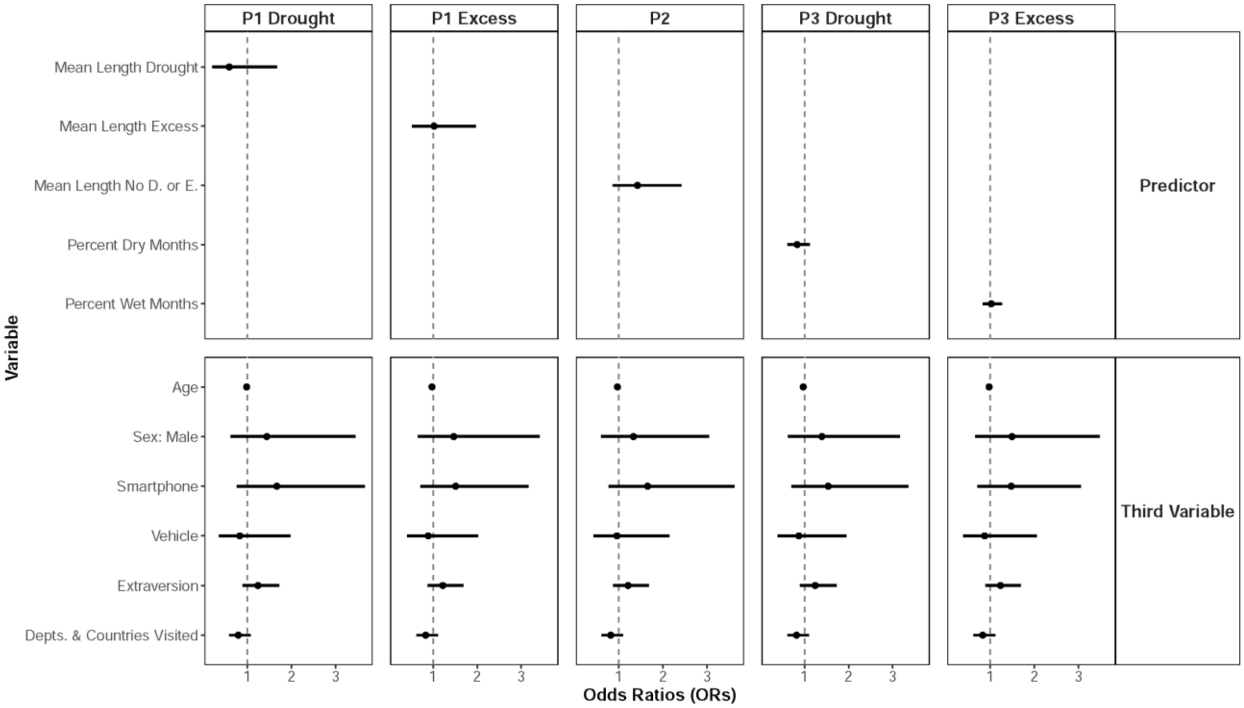
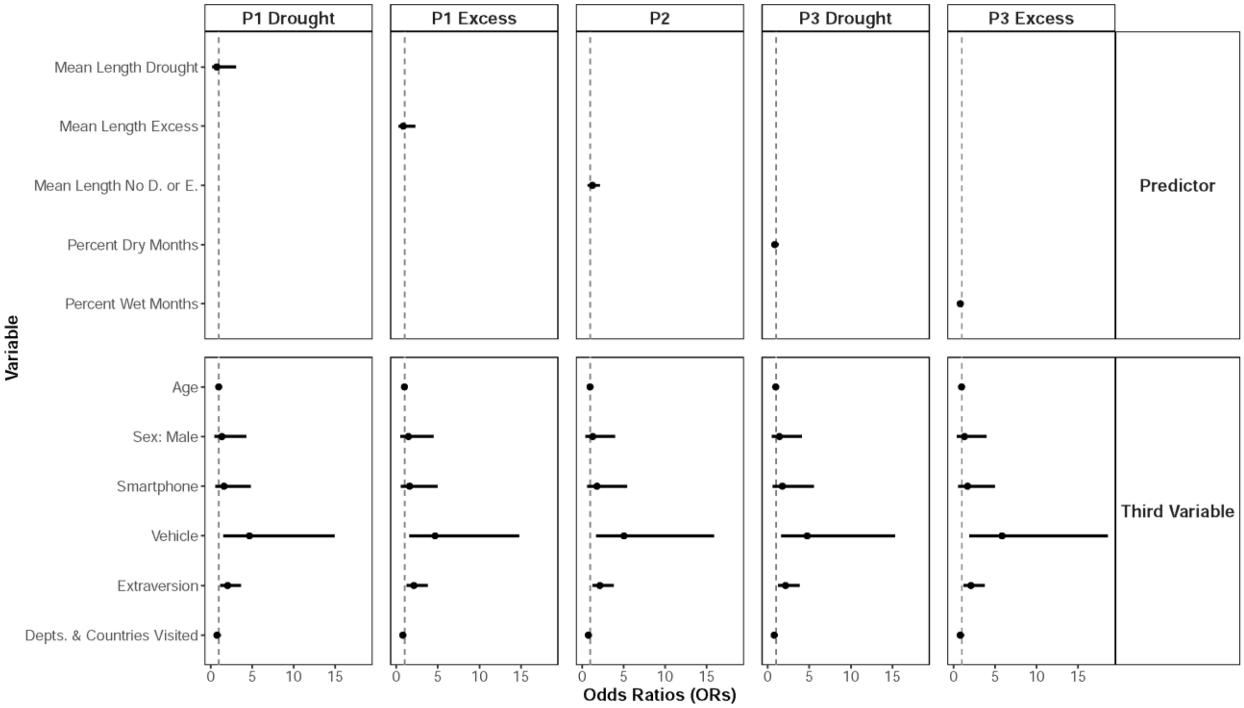


Figure S3. Parameter estimates and 90% credible intervals for a robustness check of the alternative outcome for the five main models: treating *non-consanguineal same-community relationships* as the outcome.



REFERENCES

Bürkner, P.-C., & Vuorre, M. (2019). Ordinal Regression Models in Psychology: A Tutorial. *Advances in Methods and Practices in Psychological Science*, *2*(1), 77–101. https://doi.org/10.1177/2515245918823199

Martinez-Villalobos, C., & Neelin, J. D. (2019). Why do precipitation intensities tend to follow gamma distributions? *Journal of the Atmospheric Sciences*, *76*(11), 3611–3631. https://doi.org/10.1175/JAS-D-18-0343.1

McElreath, R. (2020). *Statistical Rethinking: A Bayesian Course with Examples in R and STAN*. Chapman and Hall/CRC.

Mishra, A. K., & Singh, V. P. (2010). A review of drought concepts. *Journal of Hydrology*, *391*(1–2), 202–216. https://doi.org/10.1016/j.jhydrol.2010.07.012

Orlowsky, B., & Seneviratne, S. I. (2013). Elusive drought: Uncertainty in observed trends and short-and long-term CMIP5 projections. *Hydrology and Earth System Sciences*, *17*(5), 1765–1781. https://doi.org/10.5194/hess-17-1765-2013

Standardardized Precipitation Index (SPI) for global land surface (1949-2012). (2013). Research Data Archive at the National Center for Atmospheric Research Computational and Information Systems Laboratory. https://doi.org/10.5065/D6086397

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Multivariate Imputation by Chained Equations. *Journal Of Statistical Software*, *45*(3), 1–67. https://doi.org/10.1177/0962280206074463

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*, 1413–1432. https://doi.org/10.1007/s11222-016-9696-4