Cambridge Default Journal (2023), **XX**, 1–8

CAMBRIDGE
UNIVERSITY PRESS

ARTICLE TYPE

Supplementary Material to : "A data science approach to climate change risk assessment applied to pluvial flood occurrences for the United States and Canada"

Mathilde Bourget,[†] Mathieu Boudreault,[*‡] David A. Carozza,[‡] Jérémie Boudreault,[¶] and Sébastien Raymond[¶]

†Department of Mathematics, Université du Québec à Montréal, Montréal, QC, Canada and Collège Jean-de-Brébeuf, Montréal, QC, Canada
‡Department of Mathematics, Université du Québec à Montréal, Montréal, QC, Canada
¶Climatic Hazards and Advanced Risk Modelling, Co-operators General Insurance Company, Québec, QC, Canada and Centre Eau Terre Environnement, Institut national de la recherche scientifique, Québec, QC, Canada
*Corresponding author. Email: boudreault.mathieu@uqam.ca

This document presents additional details about statistical and machine learning methods, a bias analysis of the CRCM and more extensive tables for the portfolio applications of Section 5.

1. Statistical and machine learning methods

This section briefly describes Generalized Linear Models (GLM), Generalized Additive Models (GAM) and Random Forests (RF). We then provide more details about the implementation of these methods as well as summarized outputs for two occurrence models.

1.1 Introduction

Let Y be a random variable whose behavior we wish to explain with a set of p predictors denoted by $\mathbf{X} = [X_1, X_2, ..., X_p]$. When $Y \in \mathbb{R}$, then we can use what is known as a multiple (linear) regression to explain Y as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon$$

where ϵ is normally distributed with zero mean and the βs are coefficients of the regression. We get that

$$\mathrm{E}[Y|\mathbf{X}] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$$

and the conditional (on predictors) expectation of Y is a linear and additive function of $\mathbf{X}$.

Whenever $Y \in \{0, 1, 2, 3, ...\}$ (count response) or simply $Y \in \{0, 1\}$ (binary response), then multiple regression will not work and we need an alternative approach. For flood occurrences, we need a "regression" model that handles binary responses for classification problems. GLM provide such model for count or binary responses for example.

Now let $Y \in \{0, 1\}$. In a GLM, we represent the conditional expectation of Y as a transformation of a linear function of the predictors. Mathematically, we have

$$\mathrm{E}[Y|\mathbf{X}] = g^{-1}\left(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p\right)$$

where g is known as the link function. With $Y \in \{0, 1\}$, then $\mathrm{E}[Y|\mathbf{X}] \in [0, 1]$ is simply a probability. Therefore g^{-1} transforms a real value into $[0, 1]$. Popular functions include the logit function

$$g(x) = \log\left(\frac{x}{1-x}\right)$$

and the probit function

$$g(x) = \Phi^{-1}(x)$$

where Φ^{-1} is the quantile function of a standard normal distribution and $x \in [0, 1]$. A GLM on binary responses using the logit link function is known as a logistic regression.

A GAM extends the GLM and introduces non-linear functions of the predictors. That is we have

$$\mathrm{E}[Y|\mathbf{X}] = g^{-1}\left(\alpha + f_1(X_1) + f_2(X_2) + ... + f_p(X_p)\right)$$

with f functions to be estimated (for a predetermined functional form). We typically use splines for one or many predictors to allow for a flexible representation of non-linearities.

RF is a machine learning method based on regression or decision trees and can be used in regression or classification problems. In a process called bagging, RF combine decision trees by randomly sampling subsets of observations and predictors. The result is a multidimensional and empirical relationship between the response and the predictors. In the case of classification problems, RF outputs the predicted probability as the average probability across all underlying decision trees.

1.2 Implementation

GLM were fitted using the `glm` function included in the `stats` package loaded by default in `R`. The family, which refers to the error distribution, was set to binomial since we are working with binary responses. The link function was then set to logit to use a logistic regression. By default the `glm` function uses the iteratively reweighted least squares method to find parameters.

GAM were fitted using the `bam` function included in the `mgcv` package in `R` (Wood 2017). Just like the GLM, the family was also set to binomial with the logit link function to work with an extended logistic regression. Parameters were found using restricted maximum likelihood to avoid undersmoothing which is the default setting. We used cubic regression splines (using the `s` function of the `mgcv` package) for non-factorial variables.

RF were fitted using the `ranger` function of the `ranger` package in `R` (Wright, Wager, and Probst 2020). The number of trees was set to 500 to limit the computation time and still obtain a good prediction. The depth of each tree was set to 0, which indicates unlimited depth. Typically, the number of predictors available at each split for a classification problem is set to the rounded down square root of the number of predictors, which is equal to 3 for this work. The split rule was set to the Gini index since this is a classification problem. Long computation times (see below) prevented us from running a more systematic test of hyperparameters. Different hyperparameter sets were randomly tested with no significant change to the results.

Calibration/training of each GLM or GAM model can be achieved within seconds but training of RF models takes much longer and depends on sample size. Training a RF with the entire dataset typically takes more than an hour, but when we undersample zeroes the sample size is much smaller and the time to train a RF drops to a few minutes. For all models, predictions can be accomplished within seconds. Computation times are based on a desktop computer with a single 8-core CPU with 32 GB of RAM. Training of RF models is fully parallelized with the `ranger` package.

1.3 Summarized outputs

We provide in this section summarized outputs for key variables in the GLM and GAM models fitted with the smaller set of covariates, with undersampling (90-10) and logged population. Table 1 shows the estimate and the z-value for these variables under the GLM model (first two columns) whereas for the GAM model, we provide statistics for the smooth terms, that is the effective degrees of freedom (EDF) and the chi-square (Chi sq) value (last two columns). The z-value measures the statistical significance of a coefficient in a GLM (whether it is different from zero), the EDF

Table 1. Summarized outputs for key variables in the GLM and GAM models with the smaller set of covariates, undersampling and logged population

	GLM		GAM	
	Estimate	**z value**	**EDF**	**Chi sq**
Monthly maximum 24h precipitation	0.042	216.91	8.34	51861.9
Monthly average daily maximum temperature	0.019	11.46	7.45	1367.5
Log10 population density	0.679	118.66	7.68	10889.8
Wetland	-3.227	-49.53	5.58	2273.8
Cropland	0.021	0.86	8.80	200.9
Barren Lands	1.577	10.55	6.68	255.5
Water	-0.932	-10.27	5.94	293.4
Grassland and Forests	-0.010	-0.37	8.62	219.6

proxies the non-linearity of the fitted spline (a value close to 1 means approximately linear) while the chi-square statistic in a GAM measures the significance of the predictor and the smooth terms.

It is difficult to directly interpret the value of each coefficient in a logistic regression since it is tied to the logit link function. However, we observe that coefficients for precipitation, temperature and population are all positive, with precipitation and population being the two most statistically significant covariates. Therefore, pluvial flood probabilities increase with these three variables. For the five land use variables, the proportion of wetland is also very significant, third overall, whereas the proportion of cropland, grassland and forests is each not statistically significant.

The complete model also includes a factorial variable for the Köppen-Geiger climate classification (20), a variable for the year and dummy variables (11) for the month of the year. The Zenodo repository includes a txt file for the complete output of the GLM (and GAM) model. Looking into the complete output, we find that about half of the Köppen-Geiger climate classes are statistically significant, with an average absolute z value of about 4. We find for example there are more pluvial floods over the months of May, June, July and August, showing an average z value of nearly 20.

For the GAM, we find that the significance of the smooth terms is more important for precipitation, population, proportion of wetland and temperature respectively. In the Zenodo repository, we find similar results for the Köppen-Geiger climate classification and dummies for months.

Figure 1 shows the GAM spline functions for precipitation, temperature and population, along with 95% confidence bands (the high resolution plots are each provided on the Zenodo repository). We observe that the spline function for the monthly maximum 24h precipitation is increasing nonlinearly and plateauing for larger precipitation values. The confidence bands are very tight until the spline reaches some maximum. The reason why the spline reaches a plateau is that for very high precipitation values, there is not much of a difference between a flood occurrence probability of 99% or 99.9%. As for the temperature spline function, it is increasing only on certain intervals: below 0 degrees (Celsius) and above 30 degrees. Otherwise it is somewhat flat. Finally, population density as a proxy for urbanization clearly increases the likelihood of pluvial flooding but the relationship is very uncertain for areas with very small population density.

2. Bias analysis

To determine if the CRCM5 generates important biases in flood probabilities, one can compare *simulated* flood probabilities (with predictors computed from the CRCM5) with *predicted* flood probabilities (with predictors computed from observations). We have done such an exercise with the GLM, GAM and RF models over the common time period of 2007–2020.

Table 2 provides the distribution (over grid cells) of that difference in probabilities across the

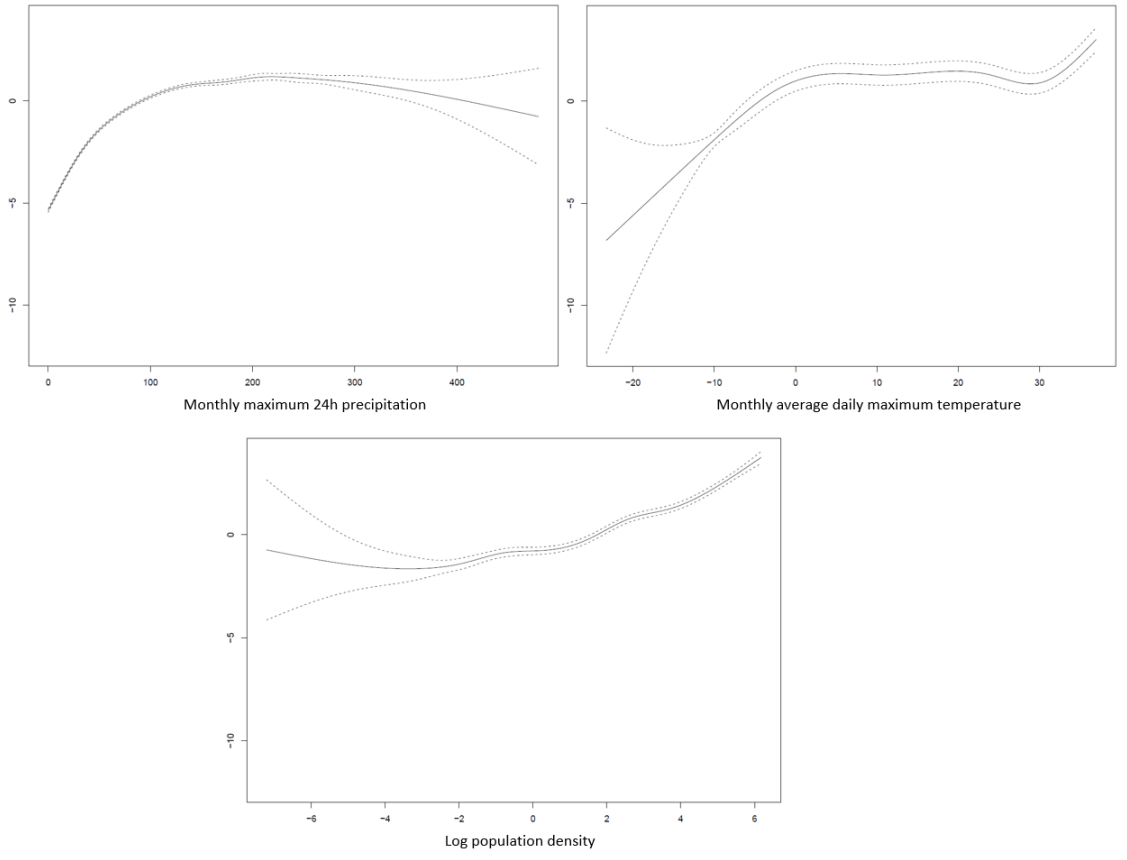

Figure 1. GAM spline functions for precipitation (top left), temperature (top right) and population (bottom) for the GAM model fitted over the United States, with the smaller set of covariates, undersampling and logged population.

Table 2. Quantiles of the difference between simulated (from the CRCM) and historical flood probabilities using the GLM, GAM and RF models over the United States and Canada

	GLM		GAM		RF	
Quantiles	USA	CAN	USA	CAN	USA	CAN
0.1%	-0.4132%	-1.2933%	-0.2331%	-0.2821%	-1.7778%	-0.0838%
1%	-0.0114%	-0.1713%	-0.0186%	-0.1026%	-0.2936%	-0.0055%
10%	0.0638%	0.0085%	0.0476%	0.0024%	0.0710%	0.0515%
25%	0.1557%	0.0224%	0.1244%	0.0210%	0.1673%	0.0769%
50%	0.4723%	0.0835%	0.3460%	0.0805%	0.3426%	0.1450%
75%	1.1368%	0.2168%	0.7756%	0.2277%	0.7077%	0.2973%
90%	1.9030%	0.4640%	1.5379%	0.4448%	1.1056%	0.5470%
99%	3.8756%	1.6976%	5.0825%	2.0039%	2.0511%	1.3404%
99.9%	6.1483%	6.1560%	11.8372%	6.0690%	3.0322%	2.4598%

United States and Canada, and over the three models. We observe that about 1% of grid cells yield negative differences, meaning about 99% of grid cells are overestimated with the CRCM5. However, the size of the overestimation remains manageable, since for the random forests, 99% of the area in the U.S. yields errors smaller than 2% (1.3% in Canada). The random forest method appears to yield smaller errors, which is consistent with its predictive capability in the test and validation sets (Sections 3.4.1 and 3.4.2). Note that there are few NAs (white cells, close to no population) in the U.S. and significantly more in Canada, which could explain why errors appear smaller in Canada.

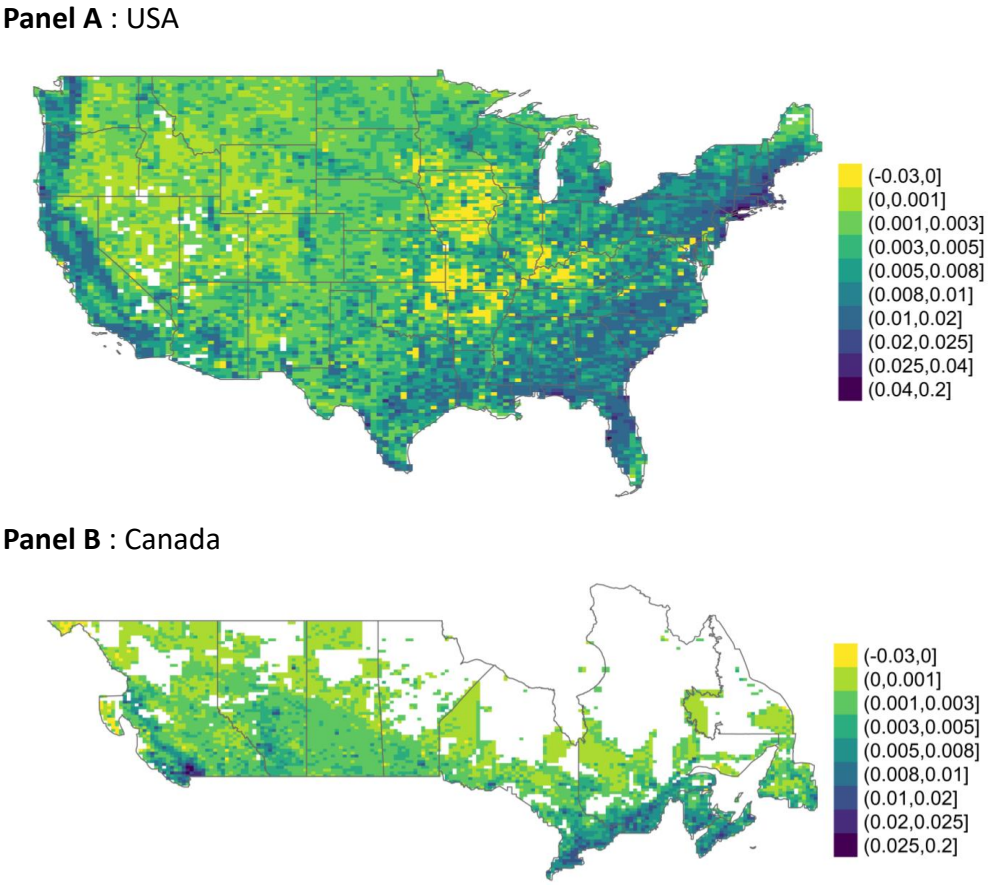

Figure 2. Difference between simulated (from the CRCM) and predicted (from observations) flood probabilities using the RF model over the United States (Panel A, top) and Canada (Panel B, bottom) and over 2007-2020. Similar plots for GLM and GAM available in the SM.

We would like to determine where errors are the smallest or the largest over Canada and the U.S. As such, Figure 2 shows the difference between the simulated flood occurrence probability (from the CRCM5) and predicted flood occurrence probability (from observations), for each grid cell, averaged over months, for the United States (Panel A, top) and Canada (Panel B, bottom) for the random forests. Similar plots for the GLM and GAM are provided in the SM. We see that errors are in general small almost everywhere, being the largest in the greater New York and Vancouver areas.

In the U.S. for example, errors are still within 1% in most key areas, that is the entire West Coast, Southern and North Eastern U.S., and within 0.3% elsewhere, namely Central U.S. In Canada, errors are the largest in South Western BC, and Southern Ontario and Quebec, in addition to New Brunswick and Nova Scotia. Overall in both countries, errors are larger in urbanized areas because their flood probabilities are larger as well.

3. Portfolio applications

This section presents the equivalent of Table 3 for the 10 Canadian provinces (Table 3) and the 10 most populous U.S. states (Table 4).

Table 3. Portfolio loss statistics for Canadian provinces and three scenarios for changes in hazard and exposure (in millions of 2020 dollars). Relative difference in % shown between parentheses (compared to the baseline scenario).

	Hazard	Exposure	Average	Std. dev.	90th perc.	95th perc.	99th perc.
	2020	2020	26	41	64	96	199
NB	2050	2020	38 (46%)	48 (18%)	89 (39%)	127 (33%)	237 (19%)
	2050	2050	43 (68%)	47 (14%)	97 (53%)	134 (40%)	223 (12%)
	2020	2020	3	12	7	17	60
PEI	2050	2020	5 (59%)	16 (24%)	13 (78%)	26 (54%)	73 (23%)
	2050	2050	6 (88%)	17 (33%)	17 (125%)	31 (83%)	79 (33%)
	2020	2020	43	84	94	163	456
NS	2050	2020	61 (42%)	97 (15%)	135 (43%)	216 (32%)	525 (15%)
	2050	2050	68 (59%)	93 (11%)	144 (52%)	218 (33%)	507 (11%)
	2020	2020	11	26	24	46	136
NL	2050	2020	16 (47%)	30 (17%)	35 (49%)	61 (32%)	156 (15%)
	2050	2050	19 (74%)	33 (27%)	42 (79%)	72 (57%)	170 (26%)
	2020	2020	26	41	64	96	199
MB	2050	2020	38 (46%)	48 (18%)	89 (39%)	127 (33%)	237 (19%)
	2050	2050	46 (78%)	49 (21%)	103 (61%)	142 (48%)	236 (18%)
	2020	2020	21	52	48	92	289
SK	2050	2020	31 (48%)	62 (18%)	75 (54%)	127 (38%)	316 (9%)
	2050	2050	38 (80%)	67 (28%)	94 (95%)	162 (75%)	341 (18%)
	2020	2020	141	273	354	581	1340
AB	2050	2020	199 (41%)	316 (16%)	493 (39%)	768 (32%)	1576 (18%)
	2050	2050	246 (75%)	352 (29%)	617 (75%)	905 (56%)	1734 (29%)
	2020	2020	389	610	998	1498	2979
BC	2050	2020	477 (23%)	648 (6%)	1143 (15%)	1737 (16%)	3376 (13%)
	2050	2050	573 (47%)	656 (8%)	1292 (30%)	1853 (24%)	3431 (15%)
	2020	2020	471	653	1154	1707	3291
QC	2050	2020	692 (47%)	767 (18%)	1601 (39%)	2234 (31%)	3872 (18%)
	2050	2050	913 (94%)	995 (53%)	2067 (79%)	2947 (73%)	5070 (54%)
	2020	2020	693	827	1613	2291	4281
ON	2050	2020	1029 (49%)	987 (19%)	2243 (39%)	3059 (34%)	4827 (13%)
	2050	2050	1285 (85%)	1175 (42%)	2713 (68%)	3701 (62%)	5842 (36%)

References

Wood, S.N. 2017. *Generalized Additive Models: An Introduction with R*. 2nd ed. Chapman / Hall/CRC.

Table 4. Portfolio loss statistics for the 10 most populous U.S. states and three scenarios for changes in hazard and exposure (in millions of 2020 dollars). Relative difference in % shown between parentheses (compared to the baseline scenario).

	Hazard	Exposure	Average	Std. dev.	90th perc.	95th perc.	99th perc.
	2020	2020	2530	1968	5175	6469	9577
CA	2050	2020	3329 (32%)	2162 (10%)	6200 (20%)	7541 (17%)	10491 (10%)
	2050	2050	4116 (63%)	2858 (45%)	7873 (52%)	9709 (50%)	13664 (43%)
	2020	2020	3208	1556	5261	6130	7825
TX	2050	2020	3886 (21%)	1703 (9%)	6140 (17%)	7004 (14%)	8870 (13%)
	2050	2050	4639 (45%)	2035 (31%)	7271 (38%)	8359 (36%)	10502 (34%)
	2020	2020	2003	1185	3581	4263	5717
FL	2050	2020	2297 (15%)	1232 (4%)	3920 (9%)	4567 (7%)	6069 (6%)
	2050	2050	2845 (42%)	1614 (36%)	4927 (38%)	5868 (38%)	8122 (42%)
	2020	2020	1815	2155	4178	5992	10896
NY	2050	2020	2142 (18%)	2240 (4%)	4616 (10%)	6514 (9%)	11306 (4%)
	2050	2050	2619 (44%)	2682 (24%)	5719 (37%)	7852 (31%)	13367 (23%)
	2020	2020	1110	677	1998	2428	3362
PA	2050	2020	1435 (29%)	740 (9%)	2422 (21%)	2894 (19%)	3877 (15%)
	2050	2050	1718 (55%)	930 (37%)	2952 (48%)	3545 (46%)	4911 (46%)
	2020	2020	1017	943	2138	2873	4615
IL	2050	2020	1341 (32%)	1064 (13%)	2673 (25%)	3404 (19%)	5119 (11%)
	2050	2050	1637 (61%)	1427 (51%)	3356 (57%)	4369 (52%)	6833 (48%)
	2020	2020	801	519	1470	1816	2555
OH	2050	2020	1098 (37%)	592 (14%)	1879 (28%)	2242 (23%)	2997 (17%)
	2050	2050	1319 (65%)	726 (40%)	2287 (56%)	2724 (50%)	3680 (44%)
	2020	2020	1066	728	1994	2454	3609
GA	2050	2020	1314 (23%)	792 (9%)	2396 (20%)	2904 (18%)	4254 (18%)
	2050	2050	1557 (46%)	971 (33%)	2881 (45%)	3521 (43%)	5281 (46%)
	2020	2020	915	534	1657	1954	2569
NC	2050	2020	1236 (35%)	616 (15%)	2062 (24%)	2402 (23%)	3112 (21%)
	2050	2050	1417 (55%)	672 (26%)	2312 (40%)	2689 (38%)	3392 (32%)
	2020	2020	405	439	917	1251	2015
MI	2050	2020	596 (47%)	523 (19%)	1240 (35%)	1624 (30%)	2487 (23%)
	2050	2050	753 (86%)	714 (63%)	1618 (76%)	2145 (71%)	3371 (67%)

Wright, Marvin N, S Wager, and P Probst. 2020. Ranger: a fast implementation of random forests. *R package version 0.12* 1.