10.1071/WF17123_AC

Supplementary Material: *International Journal of Wildland Fire*, 27(3), 164–175.

Supplementary Material

Model-specification uncertainty in future area burned by wildfires in Canada

Yan Boulanger[A,D], *Marc-André Parisien*[B] *and Xianli Wang*[C]

[A]Natural Resources Canada, Canadian Forest Service, Laurentian Forestry Centre, 1055 du P.E.P.S., PO Box 10380, Station Sainte-Foy, Québec Qc, G1V 4C7, Canada.

[B]Natural Resources Canada, Canadian Forest Service, Northern Forestry Centre, 5320 122nd Street NW, Edmonton AB, T6H 3S5, Canada.

[C]Natural Resources Canada, Canadian Forest Service, Great Lakes Forestry Centre, 1219 Queen Street East, Sault Ste Marie, ON P6A 2E5, Canada.

[D]Corresponding author. Email: yan.boulanger@canada.ca

Table S1. Description of the predictor variables (adapted from Parisien *et al*. 2014)

Variable name	Description	Units	Status*
Ignitions			
Ltg_dens	Annual density of lightning strikes per unit area, for 1995-2005	no. strikes $km^{-2}.yr^{-1}$	static
HumFoot	Human footprint, an index of human influence for the year 2005	Dimensionless	static
Vegetation			
Conif_Pct	Land cover of coniferous forest	%	static
Wetland_Pct	Land cover of wetlands	%	static
Water_Pct	Land cover of permanent water bodies	%	static
NonFuel_Pct	Landcover of nonfuel (e.g., exposed rock, open water, glaciers, recent burn)	%	static
Climate			
MaxTempWarmest	Maximum noon temperature of the warmest month	°C	dynamic
Wind90	Annual 90^{th} percentile value of wind speed	$km.h^{-1}$	dynamic

StartGrow	Start of the annual growing period (mean daily temperature of at least 5°C for five consecutive days)	Day of year	dynamic
CMI	Annual climate moisture index (precipitation minus potential evapotranspiration)	mm	dynamic
ISI99	Annual 99th percentile value of the Initial Spread Index, a CFFDRS** index of ease of fire spread	dimensionless	dynamic
FWI90	Annual 90th percentile value of the Fire Weather Index, a CFFDRS index of fire severity	dimensionless	dynamic
Topography			
SurfArea_Ratio	Ratio of surface to area, an index of topographic roughness	dimensionless	static

* Dynamic variables vary from year to year; static variables do not vary over time

** CFFDRS is the Canadian Forest Fire Danger Rating System

Correlative methods used for each modelling approach

Five different algorithms (correlative methods) were used to correlate annual area burned (burning rates) to the two predictor datasets (annual and averaged). These are further described below. All analyses were performed using R 3.3.0 (R Core Team 2016).

1) Generalized linear models (GLM)

GLM are generalizations of ordinary linear regression models for which the error distribution of the response variable can follow another distribution than the normal distribution. A link function relates the model to the response variable. In our case, we used zero-inflated negative binomial regression as GLM to link the burning rates in each hexel with the appropriate data set whereas negative binomial regressions were applied for models using the averaged data set. We use an automated genetic model selection algorithm to find the best models among all possible models that could be defined using a given predictor dataset. We allow quadratic term for *Conif_Pct* and *NonFuel_Pct* as well as first level interaction between *MaxTempWarmest* and *CMI* as well as between *MaxTempWarmest* and *StartGrowSeason* for GLM models built with the annual dataset. The same was applied for GLM models built using the averaged dataset, except for the quadratic term for *Conif_Pct* which was not included. These were in accordance with the models identified by Parisien *et al*. (2014). Best models were identified based on the AICc criterion. In order to stop the genetic algorithm, the target change in best AICc was fixed to 0.05 whereas the target change in mean AICc was set to 50 in annual models and 2 in averaged models. Maximum iteration was fixed to 5000. We used the *glmulti* function of the *glmulti* package (Calgagno 2013) in R to run the GLM analyses.

2) Multivariate Adaptive Regression Splines (MARS)

MARS is a non-parametric regression technique that makes no assumption on the relationship tying the dependent and the independent variables (Friedman, 1991). Consequently, model can handle non-linearities and complex interactions between variables. Models are built using basis functions, set as linear regressions expressed as hinge functions, that fit separate splines to distinct intervals of the predictor variables (Prasad *et al*. 2006). During first steps, pairs of basis functions are added repeatedly in order to reduce to a maximum the sum-of-square residual error. The procedure used a heuristic to select the best location of knots and variables at each step. This builds an overfitted model which is then pruned during a backward pass where least effective terms are removed sequentially. The best solution from the backward pass was found by generalized cross-validation (GCV). MARS analyses were performed using the package *earth* v4.4.9.1 (Milborrow 2017) in R.

3) Regression trees

Regression trees (RT) uses recursive binary partitioning to split the data into increasingly smaller, homogeneous subsets until a final node (leaf) is reached (Iverson and Prasad, 1998; Heikkinen *et al*. 2006). A constant is fitted within each homogeneous subset as the mean of the response variable within this very subset. Predictors and split point chosen to minimize prediction errors. The final number of splits is determined by cross-validation. RT can handle non-additive behaviour and complex interactions. The package *rpart* v4.1-10 (Therneau *et al*. 2015) in R was used to compute the RTs.

4) Gradient boosted models (GBM)

Gradient boosted modelling, or boosted regression trees, is a machine-learning algorithm that combined the regression tree approach as well as boosting. GBM combines a large

number of rather simple tree models in order to optimize predictive performance (Elith *et al*. 2008). Indeed, regression trees are fitted iteratively to the training data in such a way that the new tree will reduce a loss function. In other words, a first regression tree is fitted to the data in order to maximally reduce the loss function for a given tree size (in our case, we considered trees up to 5 levels of interaction). Then, the second tree is fitted on the residuals of the first tree. The model is then updated to include the two trees and the residuals from this latter model are calculated. These steps are repeated up to a limit fixed by the user. At each step, only a random subset of the data is used (here using 50% of the data). The final model is a linear combination of all these trees which they can be considered as terms of a regression model (Elith *et al*. 2008). Learning rates are used to shrink the contribution of each tree at each step of the model. In our case, the learning rate was fixed to 0.001 and 7950 trees were grown in the final model using the annual dataset whereas the learning rate was fixed to 0.0005 and 13350 trees were grown in the model using the averaged dataset. Interaction depth (complexity) was fixed to 5 and 10 in the annual and averaged model respectively. For both datasets, GBM building was based on a Poisson distribution. We used the *gbm* v2.1.3 package (Ridgeway 2017) in R to compute GBMs.

5) Random forests

RF is a machine-learning statistical method that fits multiple regression trees to a data set and then combines the predictions from all the trees (Cutler *et al*. 2007). First, several bootstrap samples (1500 in this study) representing about 64% of all the data are drawn from the dataset. For each bootstrap sample, a regression tree is fully grown. As oppose to a classical regression tree, the best predictor at each node is selected among a random

subset of all the predictors available. In this study, the number of variables randomly sampled as candidates at each split was set to one. Random forest was used previously by Candau and Fleming (2011) to project outbreak duration in Ontario. The *randomForest* package in R (Liaw and Wiener 2002) was used to compute RF.

References

Calcagno, V. 2013. glmulti: Model selection and multimodel inference made easy. R package version 1.0.7. https://CRAN.R-project.org/package=glmulti

Candau, J.N., Fleming, R.A. 2011 Forecasting the response of spruce budworm defoliation to climate change in Ontario. Can. J. For. Res., 41:1948–1960.

Cutler, D.R., Edwards, T.C. Jr, Beard, K.H., Cutler, A., Hess, K.T., Gibson, J., Lawler, J.J. 2007 Random forests for classification in ecology. Ecology 88:2783–2792.

Elith, A., Leathwick, J.R., Hastie, T. 2008. A working guide to boosted regression trees. J. Anim. Ecol. 77:802-813.

Friedman, J.H. 1991. Multivariate Adaptive Regression Splines. Ann. Stats 19:1-67.

Heikkinen, R.K., Luoto, M., Araújo, M.B., Virkkala, R., Thuiller, W., Sykes, M.T. 2006. Methods and uncertainties in bioclimatic envelope modelling under climate change. Progress in Physical Geography 30:751–777.

Iverson, L.R., Prasad, A.M. 1998. Predicting the abundance of 80 tree species following climate change in the eastern United States. Ecol. Monogr. 68:465-485.

Liaw, A., Wiener, M. 2002. Classification and Regression by randomForest. R News 2(3), 18-22.

Milborrow, S. 2017. earth: Multivariate Adaptive Regression Spline Models. R package version 4.4.9.1. http://CRAN.R-project.org/package=earth

Parisien, M.-A., Parks, S.A., Krawchuk, M.A., Little, J.M., Flannigan, M.D., Gowman, L.M., Moritz, M.A. 2014. An analysis of controls on fire activity in boreal Canada: comparing models built with different temporal resolutions. Ecol. Appl. 24:1341-1356.

Prasad, A.M., Iversion, L.R., Liaw, A. 2006. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. Ecosystems 9:181-199.

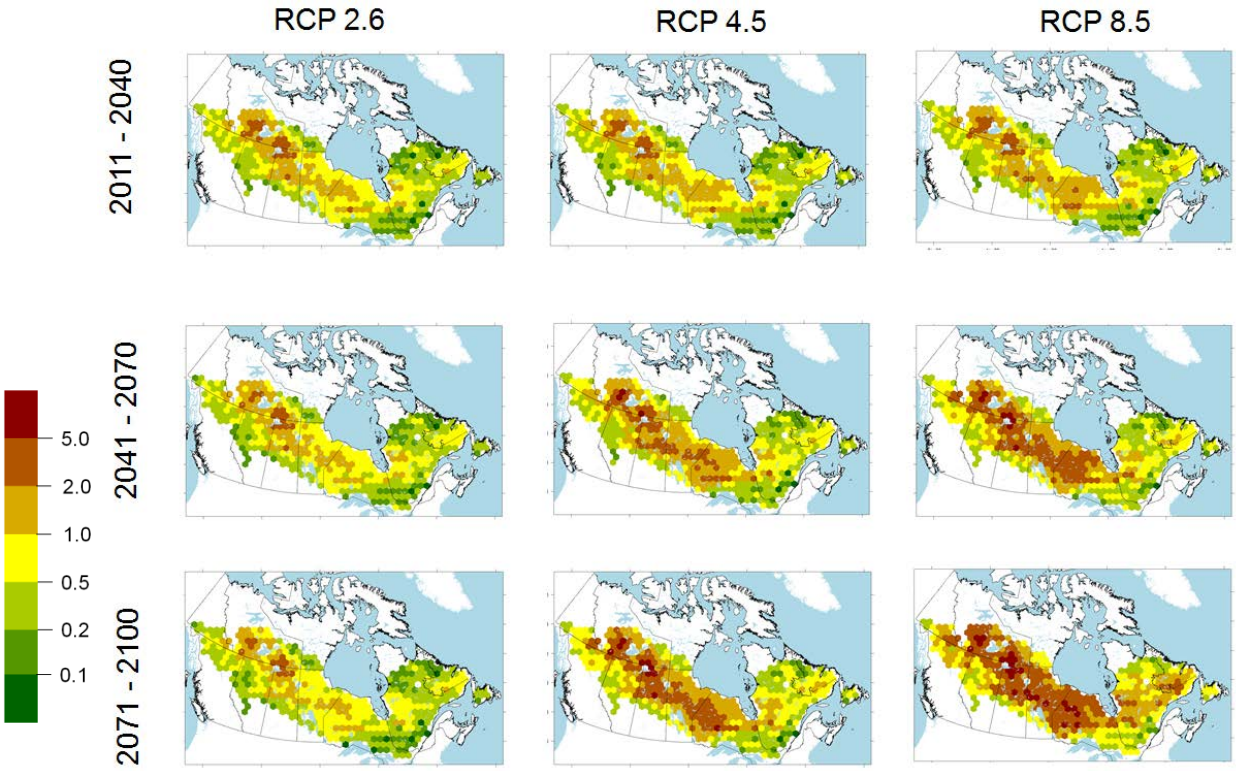
R Core Team (2016) R, A language and environment for statistical computing R Foundation for Statistical Computing, Vienna, Austria URL https,//wwwR-projectorg/.

Ridgeway, G. 2017. gbm: Generalized Boosted Regression Models. R package version 2.1.3. http://CRAN.R-project.org/package=gbm

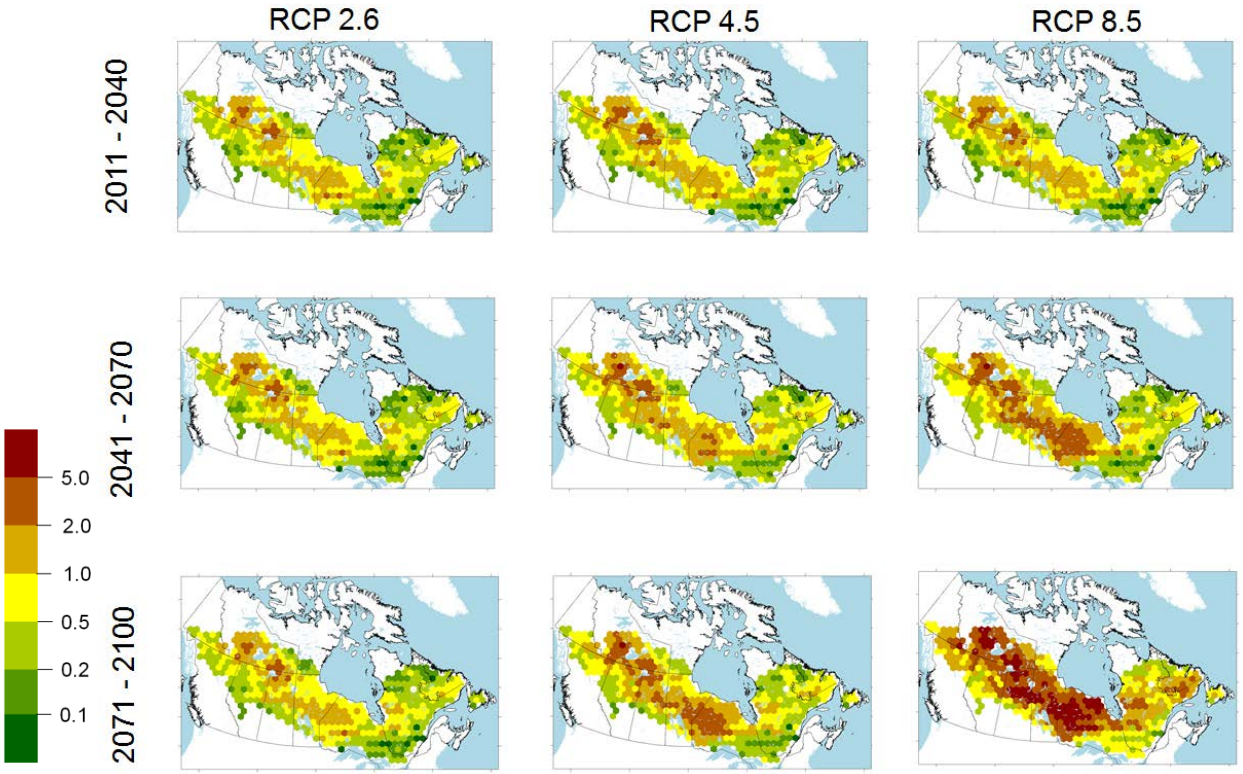
Therneau, T., Atkinson,B., Ripley, B. 2015. rpart: Recursive Partitioning and Regression Trees. R package version 4.1-10. http://CRAN.Rproject.org/package=rpart

Fig. S1. Maps of annual area burned consensus projections for the three different periods (2011 – 2040, 2041 – 2070, 2071 – 2100) under the three different GCMs (CanESM2, HadGEM, MIROC) driven by three anthropogenic climate forcing scenarios (RCP 2.6, RCP 4.5 and RCP 8.5).

a) CanESM2



b) HadGEM



c) MIROC

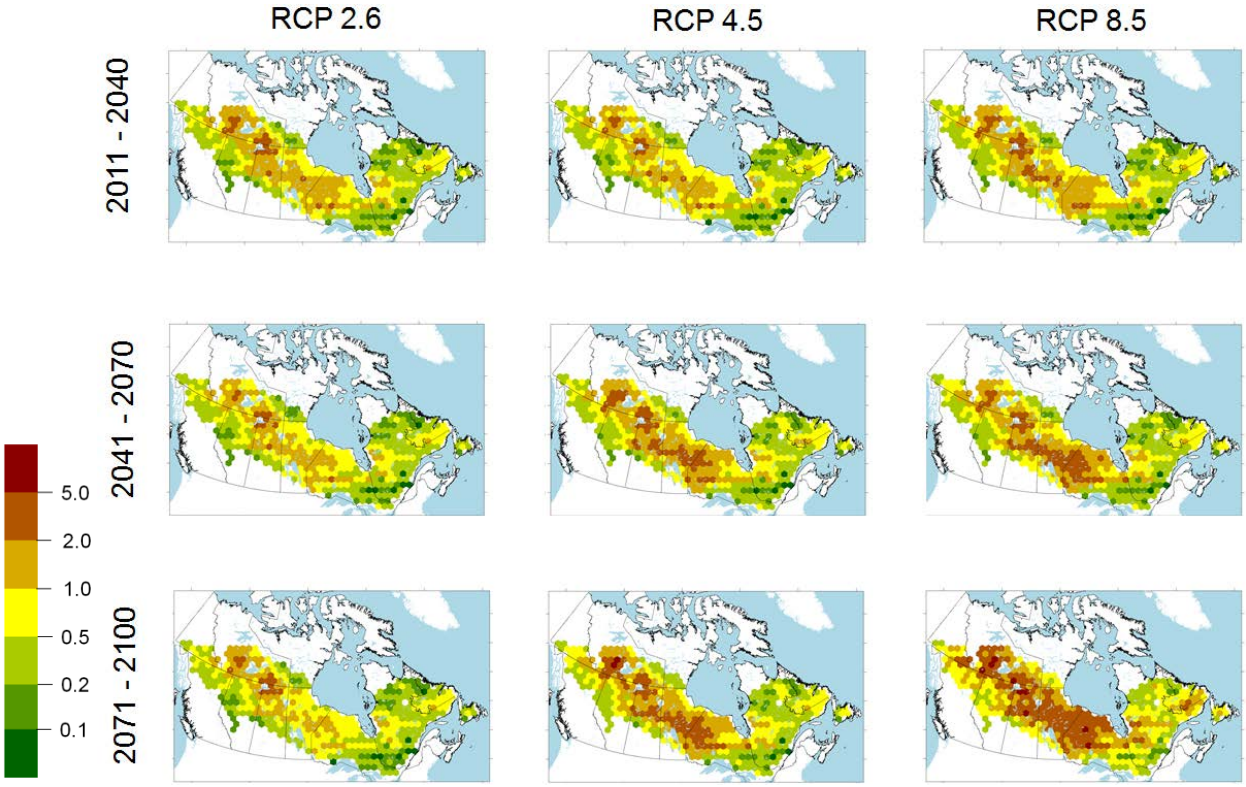
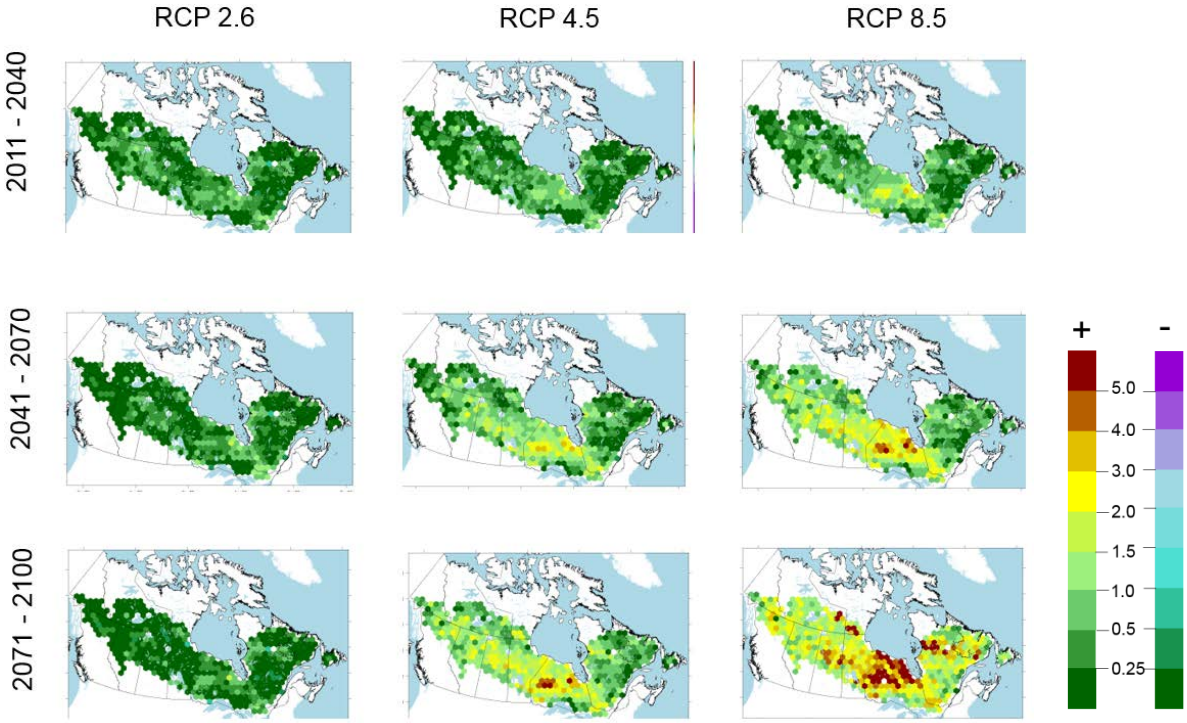
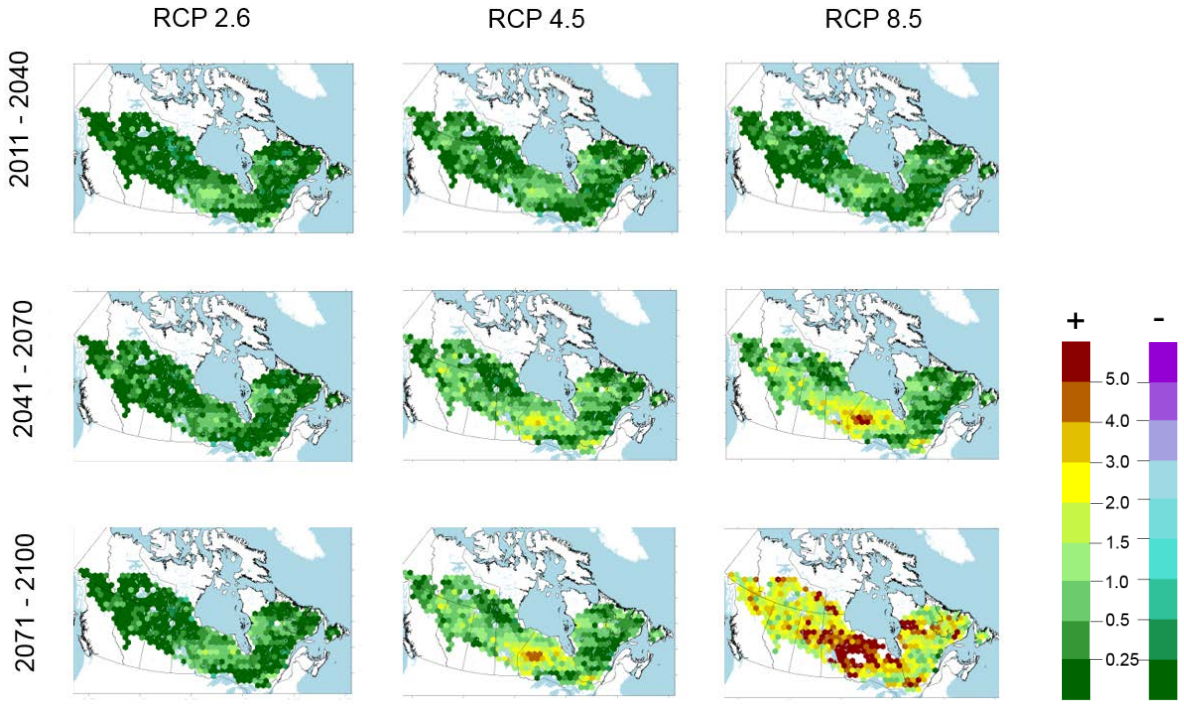


Fig. S2. Anomaly maps for annual area burned consensus projections for the three different periods (2011 – 2040, 2041 – 2070, 2071 – 2100) under the three different GCMs (CanESM2, HadGEM, MIROC) driven by three anthropogenic climate forcing scenarios (RCP 2.6, RCP 4.5 and RCP 8.5). A positive value indicates an increase in burning rates while a negative value represents a decrease in burning rates relative to the 1981 – 2010 period.

a) CanESM2



b) HadGEM



c) MIROC

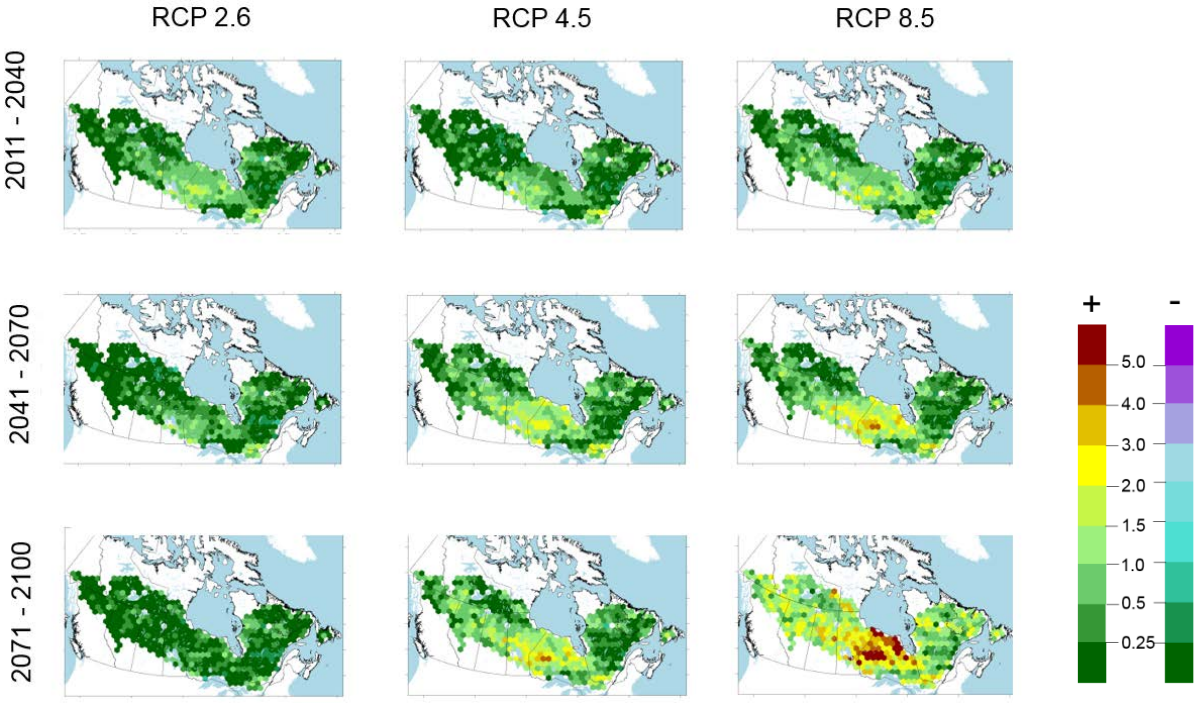
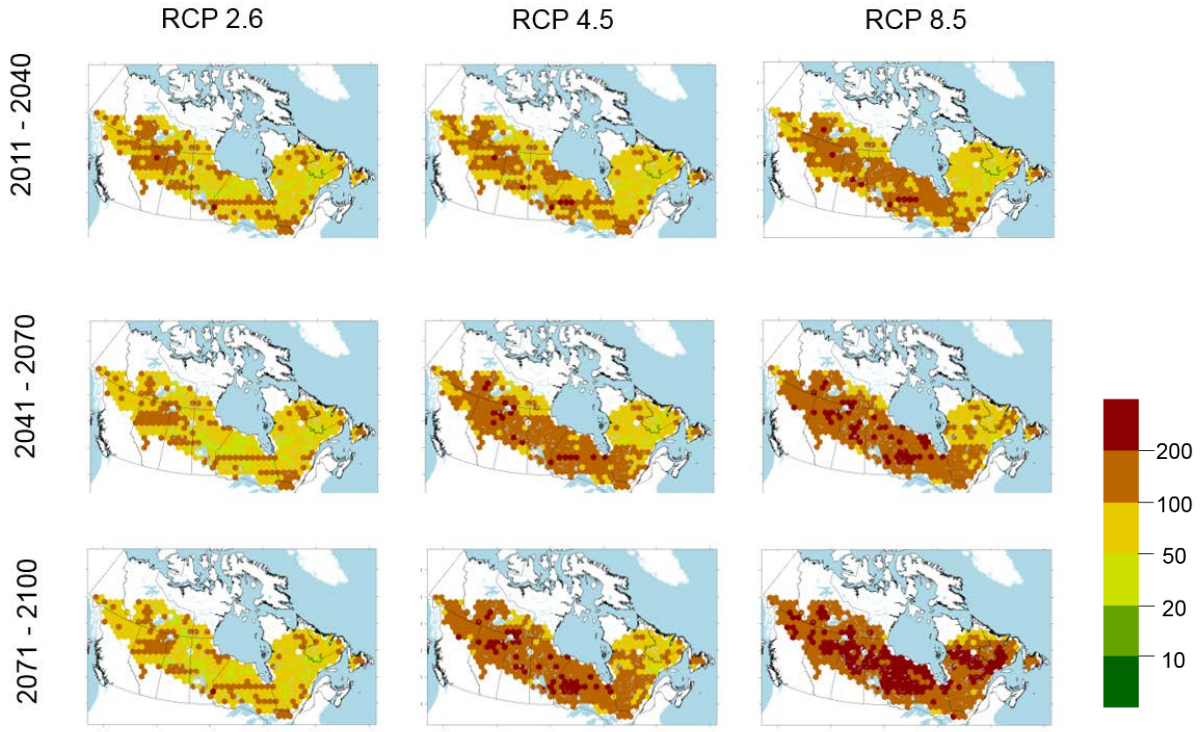
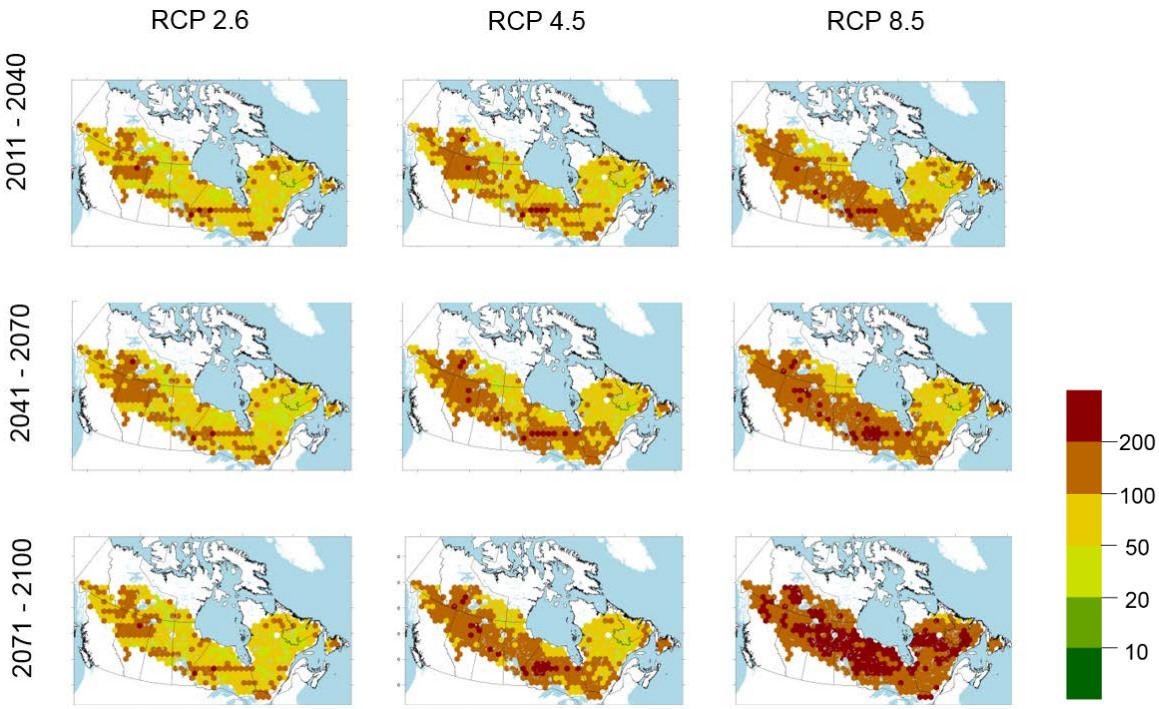


Fig. S3. Confidence interval maps for annual area burned consensus projections for the three different periods (2011 – 2040, 2041 – 2070, 2071 – 2100) under the three different GCMs (CanESM2, HadGEM, MIROC) driven by three anthropogenic climate forcing scenarios (RCP 2.6, RCP 4.5 and RCP 8.5).

a) CanESM2



b) HadGEM



c) MIROC