Supplementary Information for "SMLFire1.0: modeling wildfire activity in the western United States with stochastic machine learning"

Jatan Buch[1], A. Park Williams[2], Caroline Juang[1, 3], Winslow D. Hansen[4], and Pierre Gentine[5]

[1]Lamont-Doherty Earth Observatory, Columbia University, Palisades, NY, USA
[2]Department of Geography, University of California, Los Angeles, CA, USA
[3]Department of Earth and Environmental Sciences, Columbia University, New York, NY, USA
[4]Cary Institute of Ecosystem Studies, Millbrook, NY, USA
[5]Department of Earth and Environmental Engineering, Columbia University, New York, NY, USA

Correspondence: Jatan Buch (jb4625@columbia.edu)

Contents of this file

Divisions	Ecoregion	Level III Ecoregions	Total number of fires	Total area burned [in km^2]
Forests	Sierra Nevada	(4, 5, 9); CA	824	18428
	California (CA) North Coast	(1, 78); CA	437	17354
	CA Central Coast	(6); CA	1105	17273
	CA South Coast	(8, 85); CA	867	17111
	Pacific Northwest Mountains	(1, 4, 9, 77, 78); WA, OR	579	21673
	Northern Rockies	(15, 41)	534	11599
	Middle Rockies	(11, 16)	1947	49983
	Southern Rockies	(19, 21)	570	12961
	AZ/NM Mountains	(23, 79)	1547	33106
Deserts	AM Semidesert	(14, 81)	837	10557
	Intermountain (IM) Semidesert	(12, 18, 80)	3054	67324
	IM Desert	(13)	2248	41525
	Chihuahuan (CH) Desert	(24)	290	2557
	Columbia Plateau	(10)	793	17235
	Colorada Plateau	(20, 22)	735	6315
	Southwestern Tablelands	(26)	334	5044
Plains	Northern Great Plains	(42, 43); MT, WY	1121	17526
	High Plains	(25)	296	5498

Table S1. Summary of the Bailey's ecoregions and divisions used in our analysis. The constituent Level III (L3) ecoregions, referenced by their respective `US_L3code`, for each "Ecoregion" are outlined alongside state boundaries, wherever applicable. For example, the Northern Great Plains Ecoregion consists of three L3 ecoregions with `US_L3code`s 42 and 43 within the states of Montana (MT) and Wyoming (WY). Also shown are the total number of fires as well as total area burned (rounded upto the nearest integer) from 1984 to 2020 for each Ecoregion.

Predictor type	Identifier	Description	Timescale	Source
Climate and fire weather	VPD	Mean vapor pressure deficit	Monthly	Climgrid and PRISM
	Avg_VPDMmo	Average vapor pressure deficit in M antecedent months; $M \in \{2,3,4\}$	Monthly	Climgrid
	Tmax	Daily maximum temperature	Monthly	Climgrid
	Avg_Tmax$M$mo	Average maximum temperature in M antecedent months	Monthly	Climgrid
	Tmax_max$X$	X-day maximum temperature; $X \in \{3,7\}$	Monthly	UCLA-ERA5
	Tmin	Daily minimum temperature	Monthly	Climgrid
	Tmin_maxX	X-day minimum temperature	Monthly	UCLA-ERA5
	Prec	Precipitation total	Monthly	Climgrid
	Avg_Prec$M$mo	Average precipitation total in 3 antecedent months	Monthly	Climgrid
	$\mathrm{AntPrec}_{\mathrm{lag1}}$	Mean annual precipitation in lag year 1	Annual	Climgrid
	$\mathrm{AntPrec}_{\mathrm{lag2}}$	Mean annual precipitation in lag year 2	Annual	Climgrid
	SWE	Mean snow water equivalent	Monthly	NSIDC
	SWE_max	Daily maximum snow water equivalent	Monthly	NSIDC
	Avg_SWEMmo	Average snow water equivalent in 3 antecedent months	Monthly	NSIDC
	FM1000	1000-hour dead fuel moisture	Monthly	gridMET
	FFWI	Fosberg Fire Weather Index	Monthly	gridMET
	FFWI_max$X$	X-day maximum Fosberg Fire Weather Index	Monthly	gridMET

Predictor type	Identifier	Description	Timescale	Source
	Wind_max$X$	X-day maximum wind speed	Monthly	UCLA-ERA5
	Lightning	Lightning strike density	Monthly	NLDN
Vegetation	Forest	Fraction of forest landcover	Annual	NLCD
	Grassland	Fraction of grassland cover	Annual	NLCD
	Shrubland	Fraction of shrubland cover	Annual	NLCD
	Biomass	Aboveground biomass map	Static	Spawn et al. 2020
Human	Camp_num	Mean number of camp grounds	Static	Open source
	Camp_dist	Mean distance from nearest camp ground	Static	Open source
	Road_dist	Mean distance from nearest highway	Static	Open source
	Popdensity	Distance from nearest area with population density $> 10\,\mathrm{people/km^2}$	Annual	SILVIS
	Housedensity	Mean housing density	Annual	SILVIS
Topography	Slope	Mean slope	Static	USGS
	Southness	Mean south-facing degree of slope	Static	USGS

Table S2: Summary table of all the input predictors used in this analysis organized by the type, identifier, description, timescale, and source for each predictor. The relevant references for each source are provided in the main text.

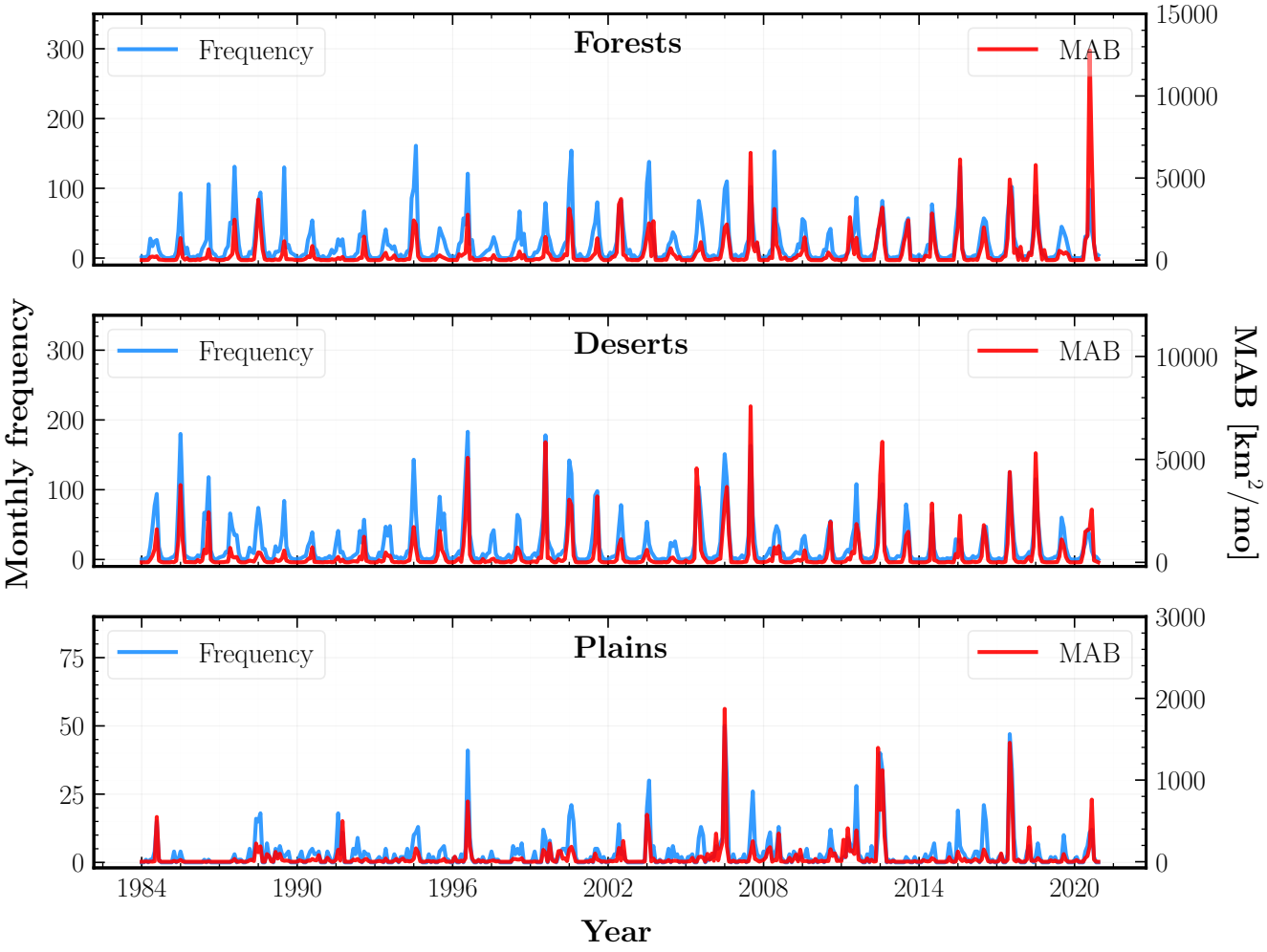

Figure S1. Observed monthly fire frequencies (blue) and monthly area burned (MAB) (red) for each of the ecological Divisions: Forests (top panel), Deserts (middle), and Plains (botom).

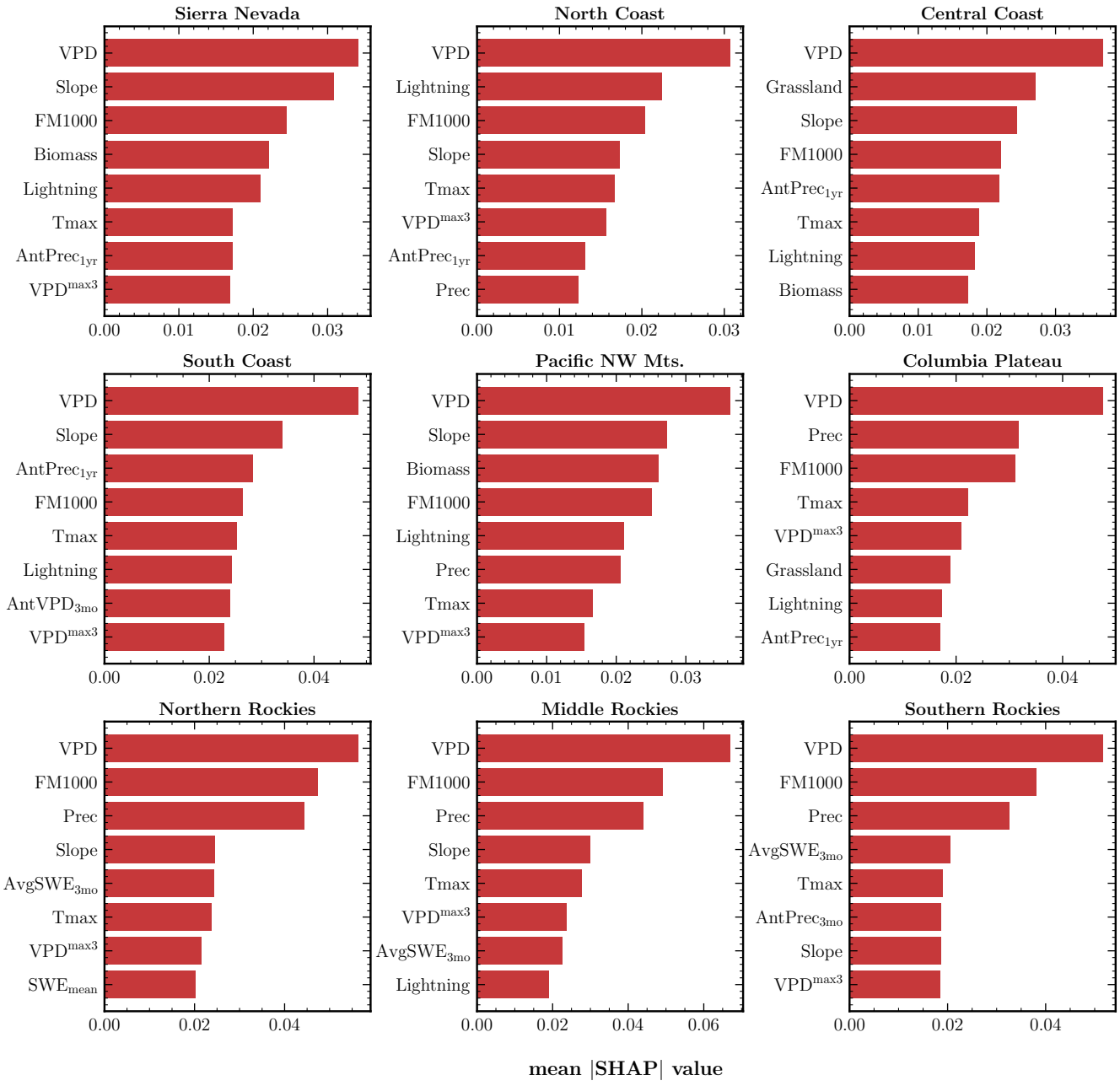

Figure S2. Mean SHAP values for the top 8 input predictors per Ecoregion of our ZIPD frequency MDN. These include all the CA Ecoregions: Sierra Nevada, North, Central, and South Coasts; Pacific NW Mountains; Columbia Plateau; and North, Middle, and Southern Rockies.

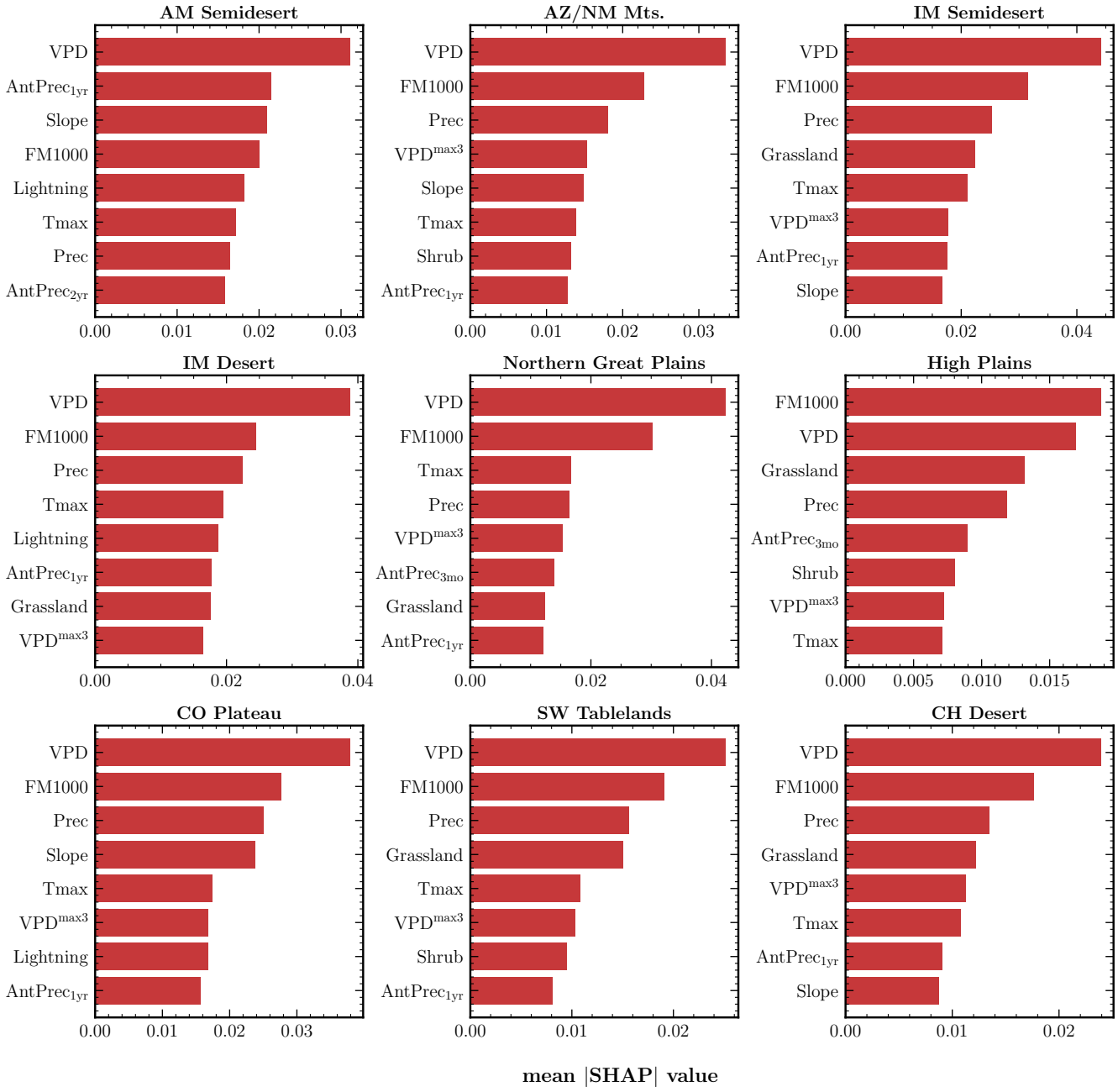

Figure S3. As in Fig. S2, but for the remaining WUS Ecoregions: American (AM) and Intermountain (IM) Semideserts, Arizona/New Mexico (AZ/NM) Mountains; Chihuahuan (CH) and IM Deserts; Northern Great and High Plains; Colorado (CO) Plateau; and Southwestern Tablelands.

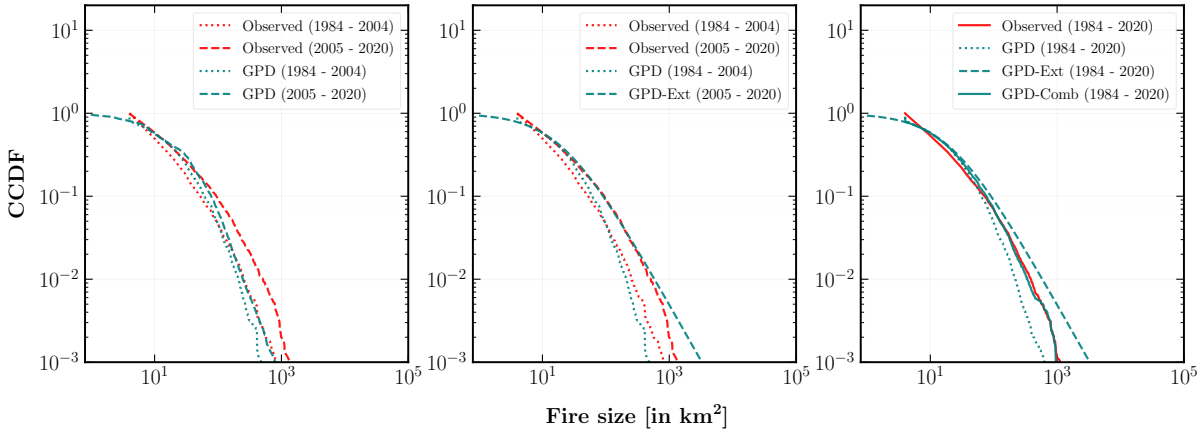

Figure S4. Complementary cumulative distribution function (CCDF) of the fire size MDN for three different cases. *Left:* CCDFs of the unweighted GPD MDN simulations (green) are plotted with those of observed (red) fire sizes ($\geq 4\,\mathrm{km}^2$) from 1984-2004 (dotted) and 2005-2020 (dashed). *Middle:* CCDFs of the unweighted GPD MDN (green, dotted) and weighted GPD (GPD-Ext) MDN simulations (green, dashed) with MDNs trained on data from 1984-2020 but plotted alongside the CCDFs of observed sizes from 1984-2004 and 2005-2020 respectively; also shown are the CCDFs for observed sizes following the legend in the previous panel. *Right:* CCDFs of the unweighted (green, dotted), weighted (green, dashed), and combined (green, solid) GPD MDN simulations alongside the CCDF of observed (red, solid) sizes from 1984-2020; the breakpoint for the combined GPD predictions is set after 2004.

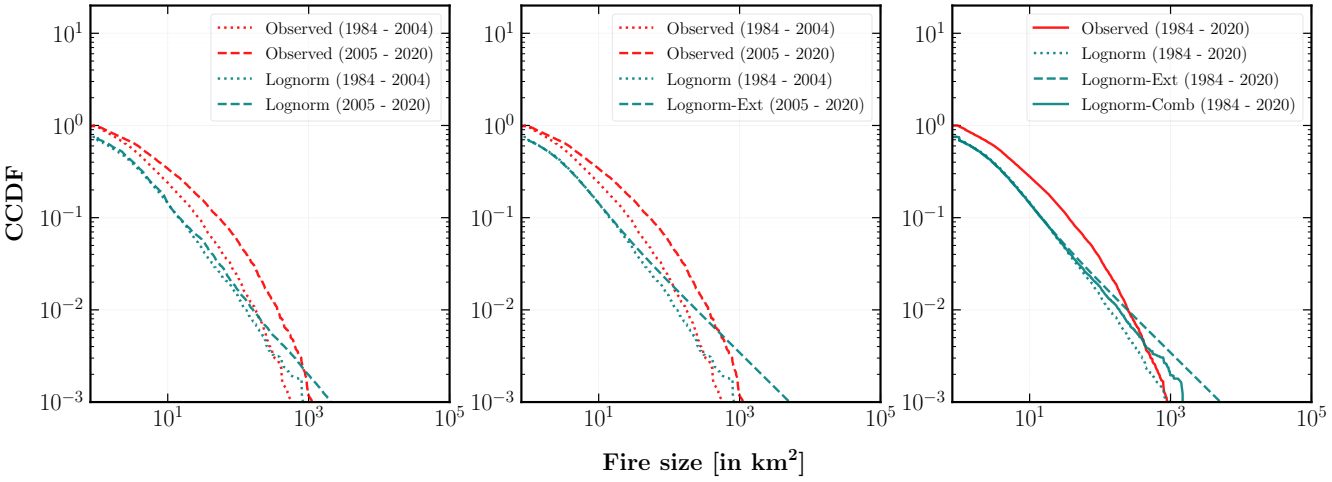

Figure S5. As in Fig. S4, but with a lognormal loss function for the MDN. Note: unlike the GPD, the lognormal distribution does not require a threshold on the fire sizes.

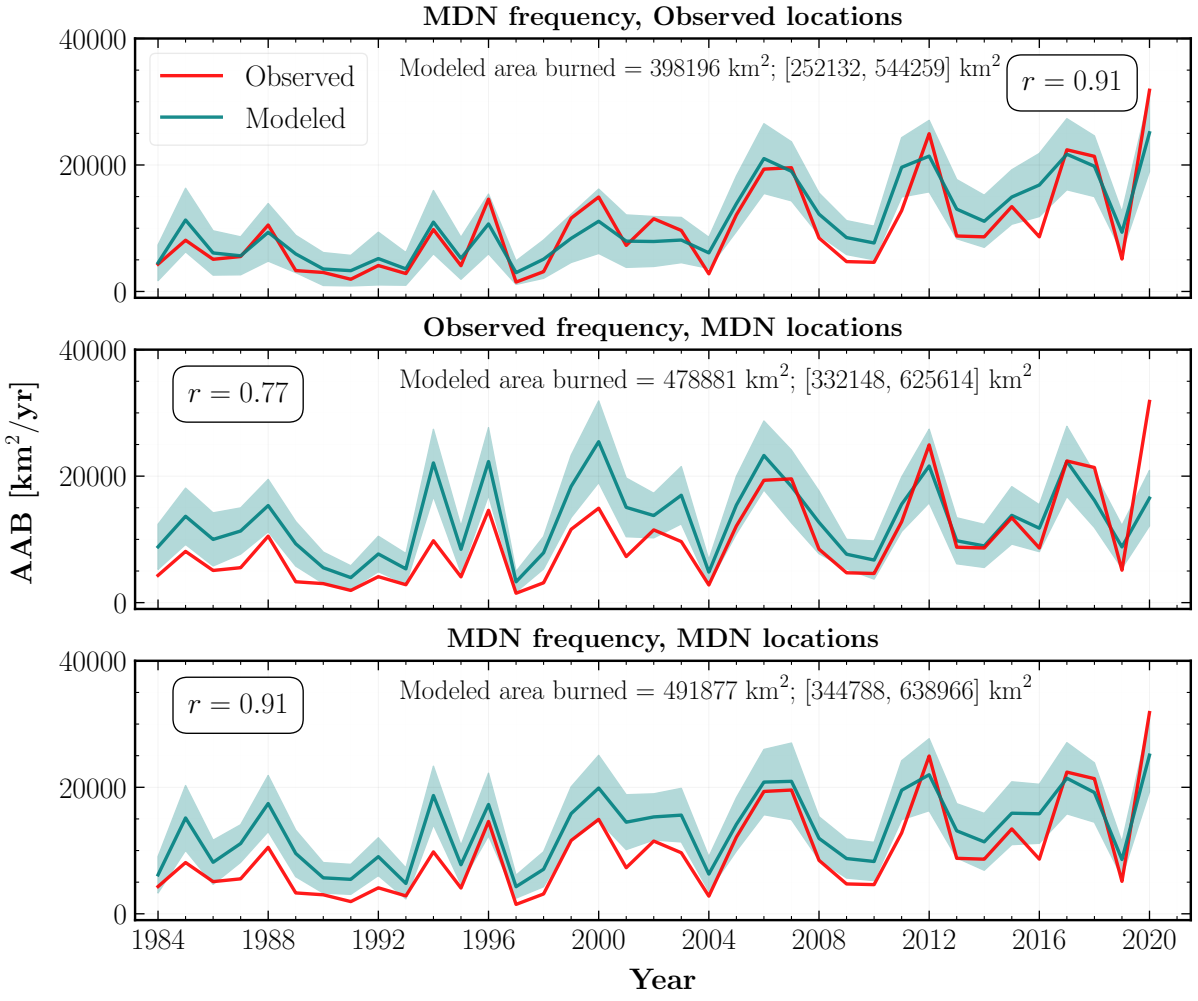

Figure S6. Cumulative observed (red) and modeled (teal) annual area burned (AAB) across the western United States from 1984 to 2020 for different combinations of fire frequencies and locations. The upper and lower panels show the AAB derived using modeled frequencies from the ZIPD MDN for each Ecoregion along with fire sizes simulated from the combined GPD model evaluated at observed and model fire locations respectively; whereas the middle panel shows the AAB computed as above except with observed frequencies and model locations. The teal shaded regions indicate 1σ uncertainty intervals for the modeled area burned aggregated over the Monte Carlo (MC) simulations of all constituent fires. The mean total area burned over the study period as well as its 1σ uncertainty interval are indicated at the top of each panel.

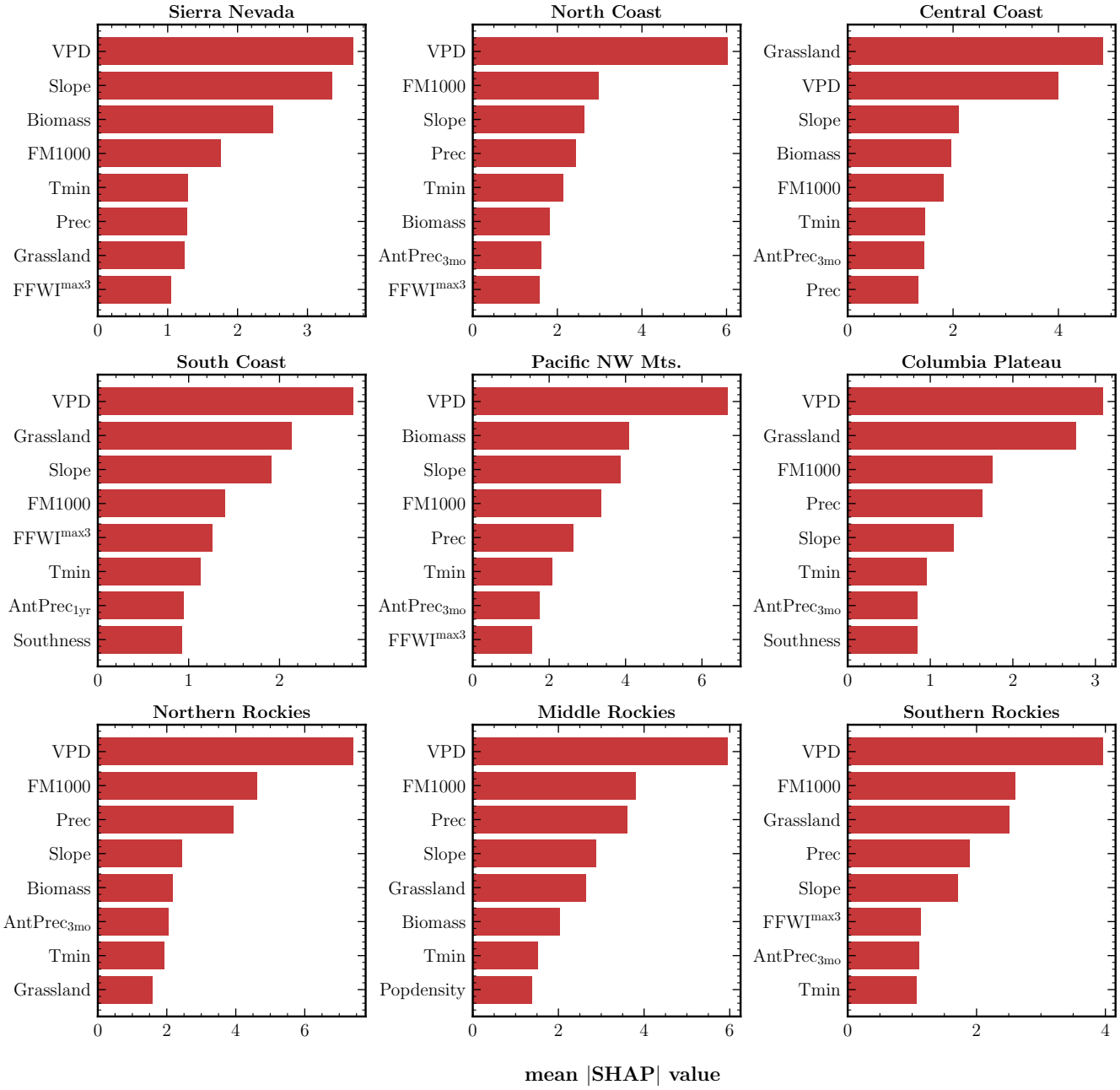

Figure S7. Mean SHAP values for the top 8 input predictors per ecoregion of the GPD size MDN. These include all the CA Ecoregions: Sierra Nevada, North, Central, and South Coasts; Pacific NW Mountains; Columbia Plateau; and North, Middle, and Southern Rockies.

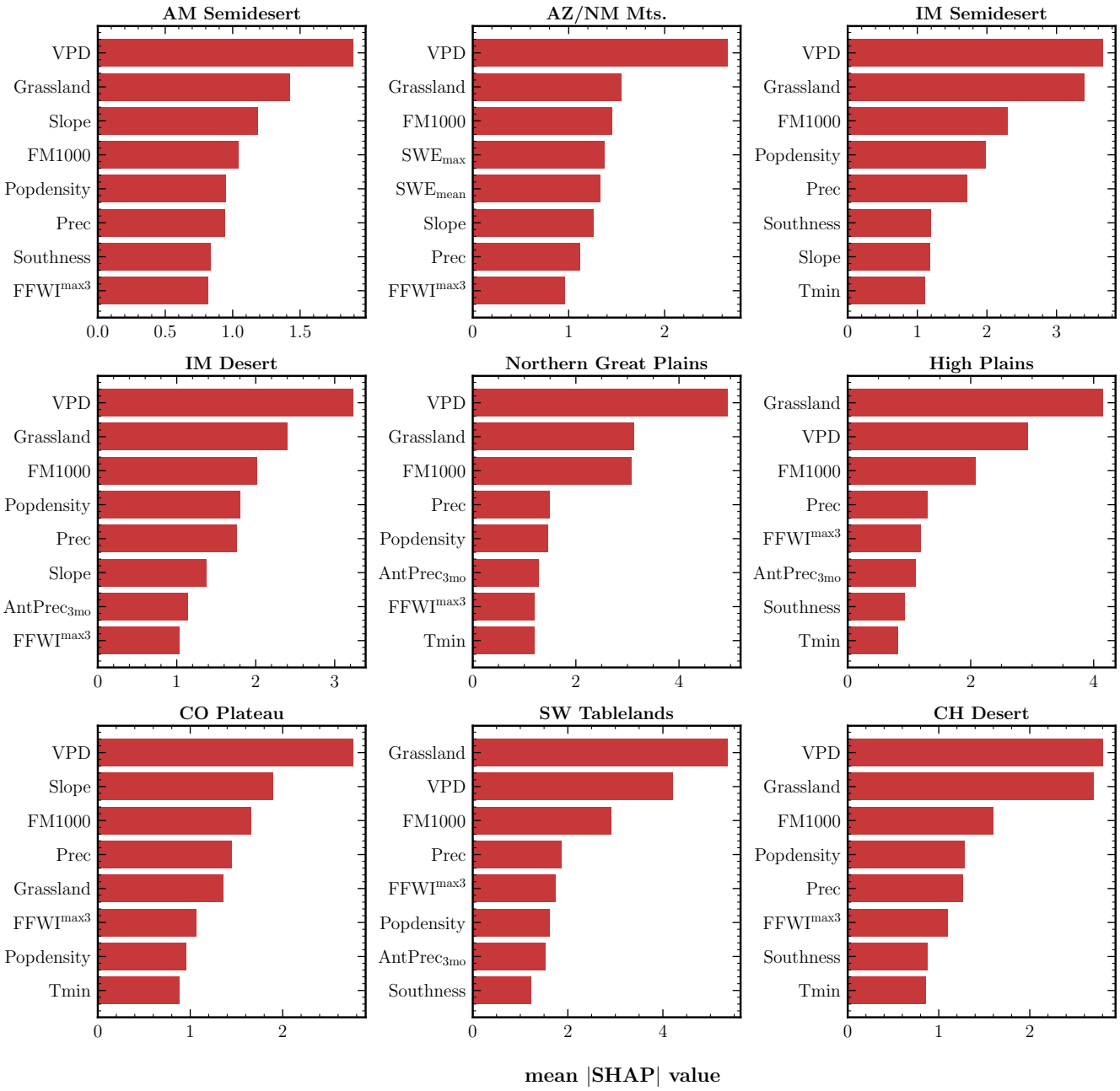

Figure S8. As in Fig. S7, but for the remaining WUS Ecoregions: American (AM) and Intermountain (IM) Semideserts, Arizona/New Mexico (AZ/NM) Mountains; Chihuahuan (CH) and IM Deserts; Northern Great and High Plains; Colorado (CO) Plateau; and Southwestern Tablelands.

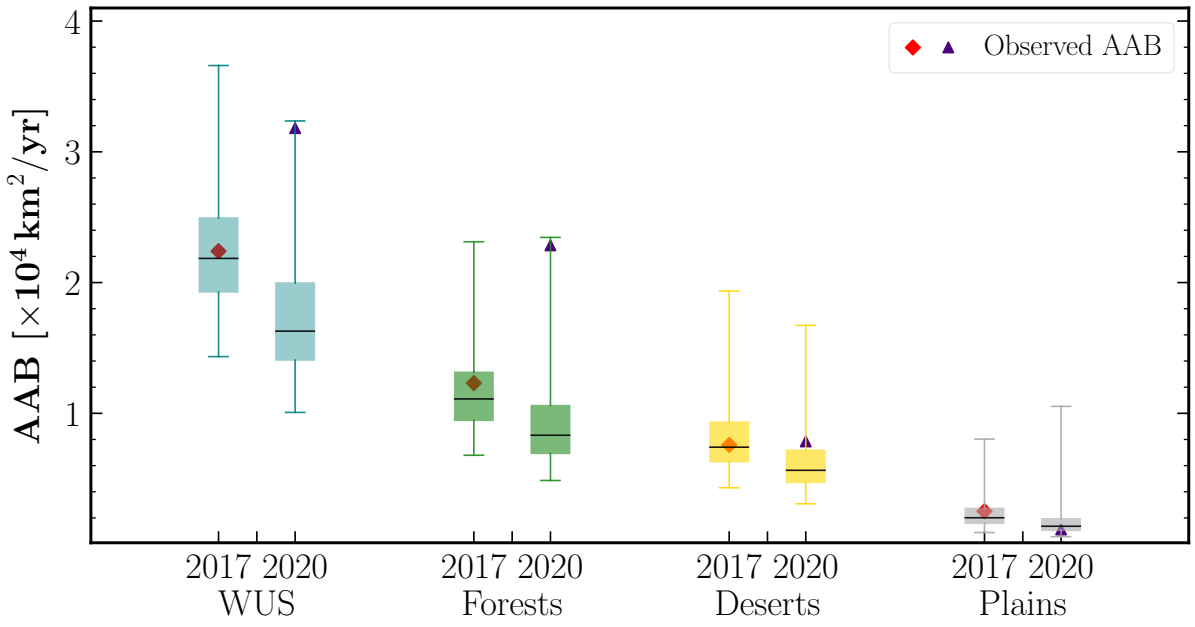

Figure S9. Boxplots of modeled annual area burned (AAB) for two extreme fire years, 2017 and 2020, for the entire western United States (WUS) (teal) and three Divisions organized by their primary vegetation types: Forests (green), Deserts (yellow), and Plains (gray). The lower and upper whiskers of each boxplot indicate the 0.5^{th} and 99.5^{th} percentile of the predicted AAB distribution, whereas the horizontal black line represents its median value. Also shown for reference are the observed AAB for both 2017 (red diamond) and 2020 (indigo traingle).

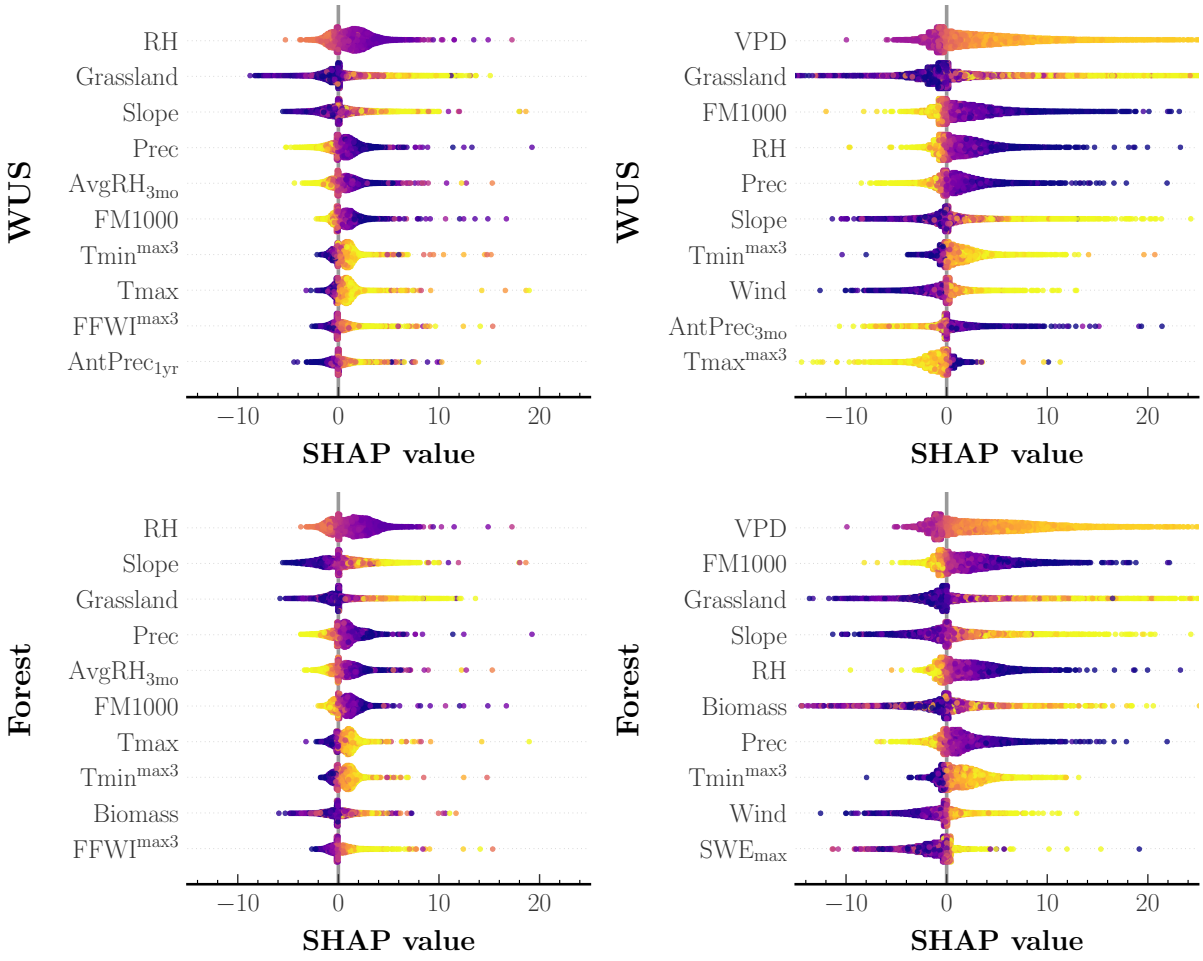

Figure S10. SHapley Additive exPlanation (SHAP) analysis of the fire size MDN model outputs for different sets of input predictors. *Left column:* SHAP summary plots with relative humidity (RH) and average RH over 3 antecedent months ($\mathrm{AvgRH}_{\mathrm{3mo}}$) predictors instead of their VPD counterparts for the entire WUS (top panel) and Forest Division (bottom). *Right column:* SHAP summary plots with both VPD and RH predictors as well as their antecedent counterparts for the entire WUS (top panel) and Forest Division (bottom). Each colored point along the x-axis represents an individual prediction with the color corresponding to high (yellow) or low (indigo) values of the respective input predictor.