Supplement of Geosci. Model Dev., 16, 3407–3433, 2023
https://doi.org/10.5194/gmd-16-3407-2023-supplement

Geoscientific Model Development Open Access EGU

Supplement of

SMLFire1.0: a stochastic machine learning (SML) model for wildfire activity in the western United States

Jatan Buch et al.

Correspondence to: Jatan Buch (jb4625@columbia.edu)

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	Resolution	Timescale	Source
Climate and fire weather	VPD	Mean vapor pressure deficit	5 km	Monthly	Climgrid and PRISM
	$\mathrm{AntVPD}_{M\mathrm{mon}}$	Average VPD in M antecedent months; $M \in \{2,3,4\}$	5 km	Monthly	Climgrid
	$\mathrm{VPD}^{\mathrm{max}X}$	Maximum X-day running average of VPD; $X \in \{3,5,7\}$	9 km	Monthly	UCLA-ERA5
	Tmax	Daily maximum temperature	5 km	Monthly	Climgrid
	$\mathrm{AntTmax}_{M\mathrm{mon}}$	Average maximum temperature in M antecedent months	5 km	Monthly	Climgrid
	$\mathrm{Tmax}^{\mathrm{max}X}$	Maximum X-day running average of Tmax	9 km	Monthly	UCLA-ERA5
	Tmin	Daily minimum temperature	5 km	Monthly	Climgrid
	$\mathrm{Tmin}^{\mathrm{max}X}$	Maximum X-day running average of Tmin	9 km	Monthly	UCLA-ERA5
	Prec	Precipitation total	5 km	Monthly	Climgrid
	$\mathrm{AntPrec}_{M\mathrm{mon}}$	Average precipitation total in M antecedent months	5 km	Monthly	Climgrid
	$\mathrm{AntPrec}_{\mathrm{lag1}}$	Mean annual precipitation in lag year 1	5 km	Annual	Climgrid
	$\mathrm{AntPrec}_{\mathrm{lag2}}$	Mean annual precipitation in lag year 2	5 km	Annual	Climgrid
	$\mathrm{SWE}_{\mathrm{mean}}$	Mean snow water equivalent	500 m	Monthly	NSIDC
	$\mathrm{SWE}_{\mathrm{max}}$	Daily maximum snow water equivalent	500 m	Monthly	NSIDC
	$\mathrm{AvgSWE}_{M\mathrm{mon}}$	Average snow water equivalent in M antecedent months	500 m	Monthly	NSIDC
	FM1000	1000-hour dead fuel moisture	4 km	Monthly	gridMET
	FFWI	Fosberg Fire Weather Index	9 km	Monthly	UCLA-ERA5

Predictor type	Identifier	Description	Resolution	Timescale	Source
	$\mathrm{FFWI}^{\mathrm{max}X}$	Maximum X-day running average of Fosberg Fire Weather Index	9 km	Monthly	UCLA-ERA5
	Wind	Monthly mean wind speed	9 km	Monthly	UCLA-ERA5
	$\mathrm{Wind}^{\mathrm{max}X}$	Maximum X-day running average of wind speed	9 km	Monthly	UCLA-ERA5
	Lightning	Lightning strike density	500 m	Monthly	NLDN
Vegetation	Forest	Fraction of forest landcover	30 m	Annual	NLCD
	Grassland	Fraction of grassland cover	30 m	Annual	NLCD
	Shrubland	Fraction of shrubland cover	30 m	Annual	NLCD
	Biomass	Aboveground biomass map	300 m	Static	Spawn et al. 2020
Human	Camp_num	Mean number of camp grounds	1km	Static	Open source
	Camp_dist	Mean distance from nearest camp ground	1km	Static	Open source
	Road_dist	Mean distance from nearest highway	1km	Static	Open source
	Popdensity	Mean population density	1km	Annual	SILVIS
	Pop10_dist	Distance from nearest area with population density $> 10\,\mathrm{people/km}^2$	1km	Annual	SILVIS
	Housedensity	Mean housing density	1km	Annual	SILVIS
Topography	Slope	Mean slope	1m	Static	USGS
	Southness	Mean south-facing degree of slope	1m	Static	USGS

Table S2: Summary table for all input predictors organized by type, identifier, description, spatial resolution of raw data, timescale, and source. All predictors are aggregated to a 12 km spatial resolution while performing the statistical analysis. Considering each predictor's M antecedent month average and maximum X-day running average components as distinct predictors, the total number of predictors adds up to 51.

Predictors	Qualitative effect		Comments
	Fire frequency	Fire size	
VPD, $\mathrm{AntVPD}_{M\mathrm{mon}}$, $\mathrm{VPD}^{\mathrm{max}X}$	↑	↑	VPD on multiple timescales, from weekly to seasonal, is positively correlated with both fire frequency and size
Tmax, $\mathrm{AntTmax}_{M\mathrm{mon}}$, $\mathrm{Tmax}^{\mathrm{max}X}$	↑	↑	Tmax on multiple timescales, from weekly to seasonal, is positively correlated with both fire frequency and size
Tmin, $\mathrm{Tmin}^{\mathrm{max}X}$	/	↑	Both extreme Tmin and monthly mean Tmin are positively correlated with fire size; Tmin is not a significant predictor for fire frequency
Prec, $\mathrm{AntPrec}_{M\mathrm{mon}}$	↓	↓	Prec on multiple timescales, from weekly to seasonal, is negatively correlated with both fire frequency and size.
$\mathrm{AntPrec}_{\mathrm{lag1}}$, $\mathrm{AntPrec}_{\mathrm{lag2}}$	↑	↑	Annual mean of Prec in lagging years, a proxy for biomass growth, is positively correlated with fire frequency and size.
$\mathrm{SWE}_{\mathrm{mean}}$, $\mathrm{AvgSWE}_{M\mathrm{mon}}$, $\mathrm{SWE}_{\mathrm{max}}$	↓	↓	Snow water equivalent on multiple timescales, from weekly to seasonal, is negatively correlated with both fire frequency and size
FM1000	↓	↓	1000-hour dead fuel moisture is negatively correlated with fire frequency and size
FFWI, $\mathrm{FFWI}^{\mathrm{max}X}$	↑	↑	Mean and extreme values of FFWI are positively correlated with fire frequency and size
$\mathrm{Wind}^{\mathrm{max}X}$	/	↑	Monthly maxima of X-day mean wind speed is positively correlated with fire size; wind speed is not a significant predictor for fire frequency
Biomass	↕	↑	Spatial variance in biomass is positively correlated with fire size, however its effect on fire frequency is ambiguous with potential confounding by human action predictors
Grassland, Shrubland	↑	↑	Fraction of grassland and shrubland cover increases fuel flammability and continuity over a landscape, and is thus positively correlated with fire frequency and fire size

Lightning	$\uparrow$	/	Increased lightning strike density contributes additional ignitions, and is positively correlated with fire frequency; Lightning is not a significant predictor of fire size
Slope	$\uparrow$	$\uparrow$	Slope is positively correlated with fire frequency and size since the rate of fire spread is proportional to the degree of slope
Southness	$\uparrow$	$\uparrow$	Southness, or mean south-facing degree of slope, dictates the level of solar insolation and is positively correlated with fire frequency and size
Pop10_dist	$\updownarrow$	$\uparrow$	Increased distance from areas with population density greater than 10 km^2 is a correlate of remoteness leading to a larger fire size. Its effect on fire frequency is ambiguous since remote areas experience fewer ignitions while also having reduced access to early fire containment efforts

Table S3: Summary of the physical meaning of important model predictors as well as their qualitative effect on fire frequency and size. The symbols $\uparrow$, $\downarrow$, $\updownarrow$, and / refer to positive, negative, ambiguous, and insignificant correlations between a predictor and fire response variable respectively.

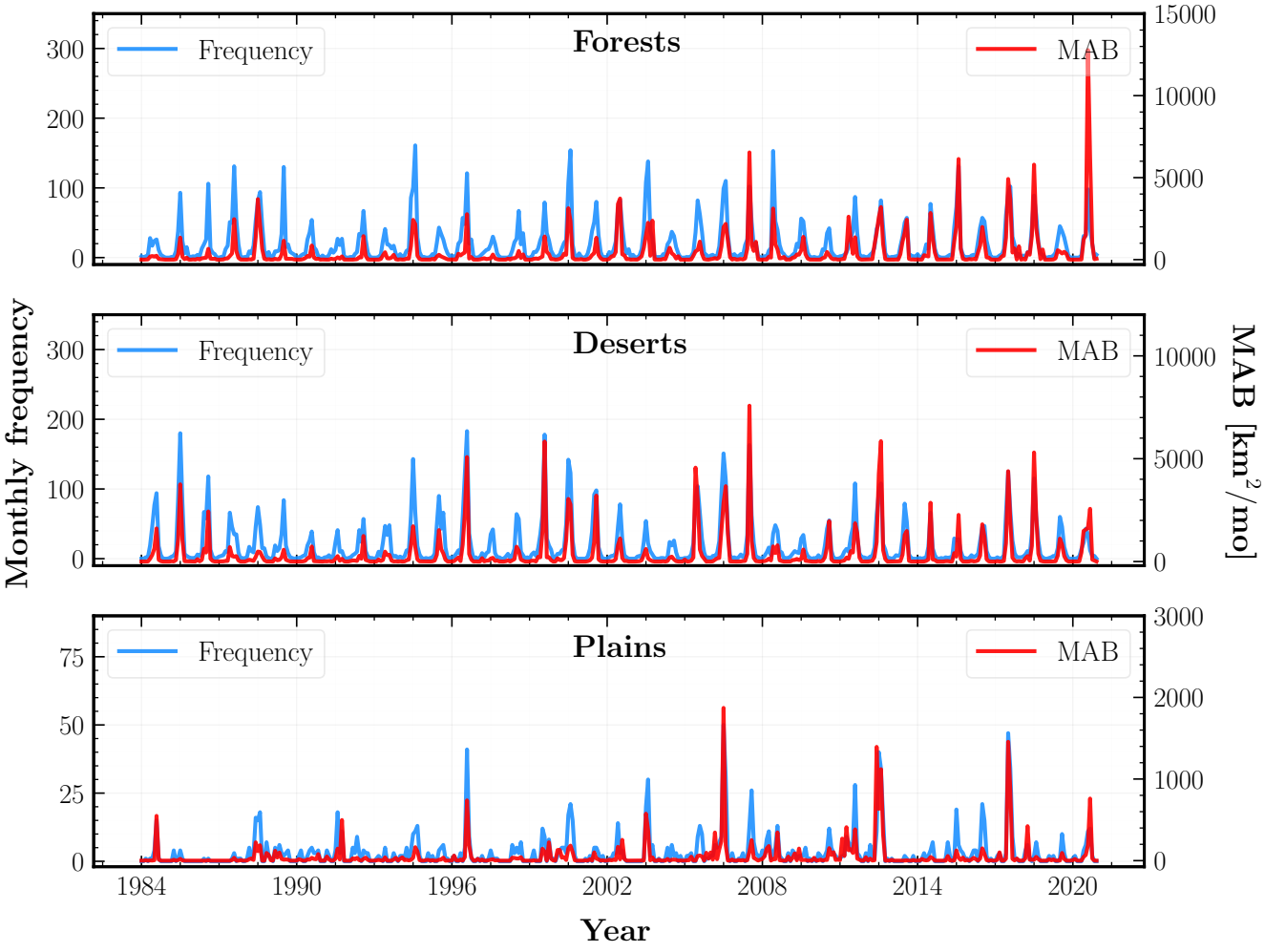

Figure S1. Observed monthly fire frequencies (blue) and monthly area burned (MAB) (red) for each of the ecological Divisions: Forests (top panel), Deserts (middle), and Plains (botom).

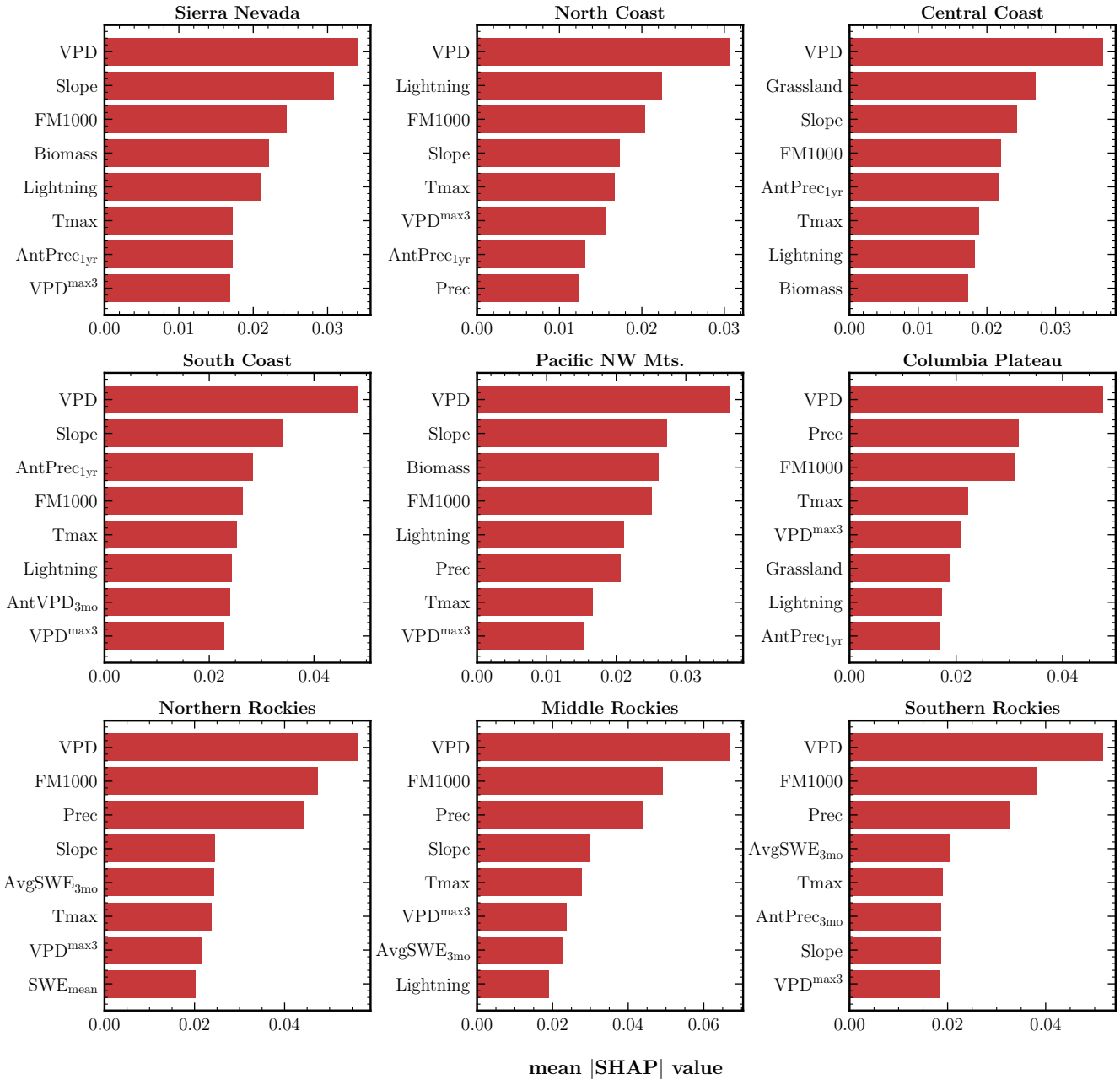

Figure S2. Mean SHAP values for the top 8 input predictors per Ecoregion of our zero-inflated Poisson distribution (ZIPD) frequency Mixture Density Network (MDN). These include all the CA Ecoregions: Sierra Nevada, North, Central, and South Coasts; Pacific NW Mountains; Columbia Plateau; and North, Middle,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

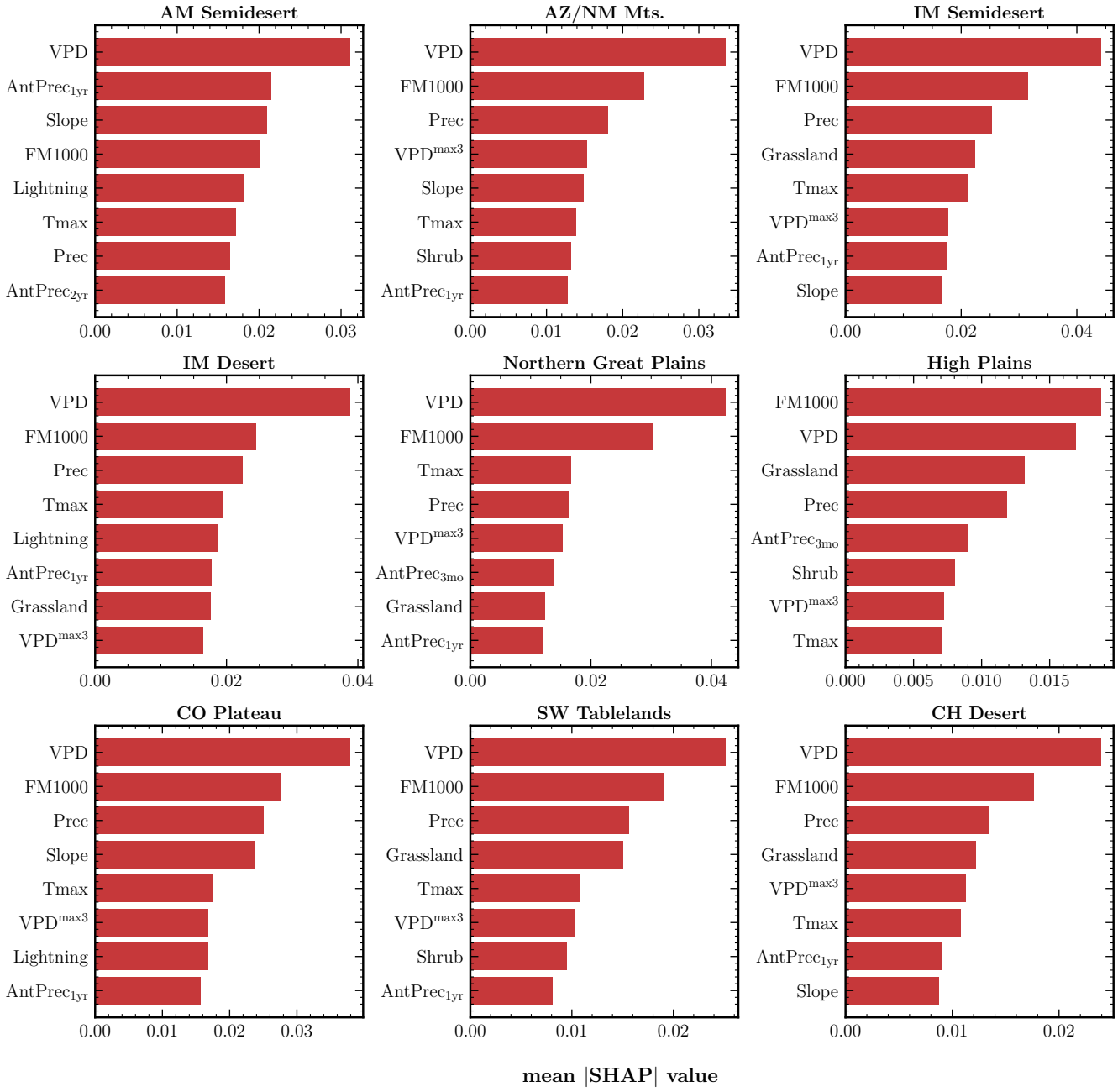

Figure S3. As in Fig. S2, but for the remaining WUS Ecoregions: American (AM) and Intermountain (IM) Semideserts, Arizona/New Mexico (AZ/NM) Mountains; Chihuahuan (CH) and IM Deserts; Northern Great and High Plains; Colorado (CO) Plateau; and Southwestern Tablelands.

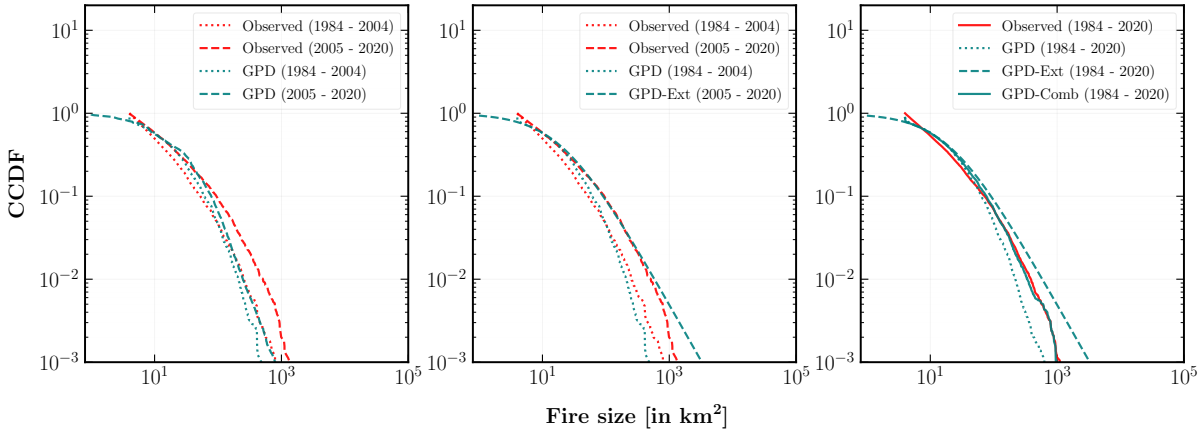

Figure S4. Complementary cumulative distribution function (CCDF) of the fire size MDN for three different cases. *Left:* CCDFs of the unweighted GPD MDN simulations (green) are plotted with those of observed (red) fire sizes ($\geq 4\,\mathrm{km}^2$) from 1984-2004 (dotted) and 2005-2020 (dashed). *Middle:* CCDFs of the unweighted Generalized Pareto distribution (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.

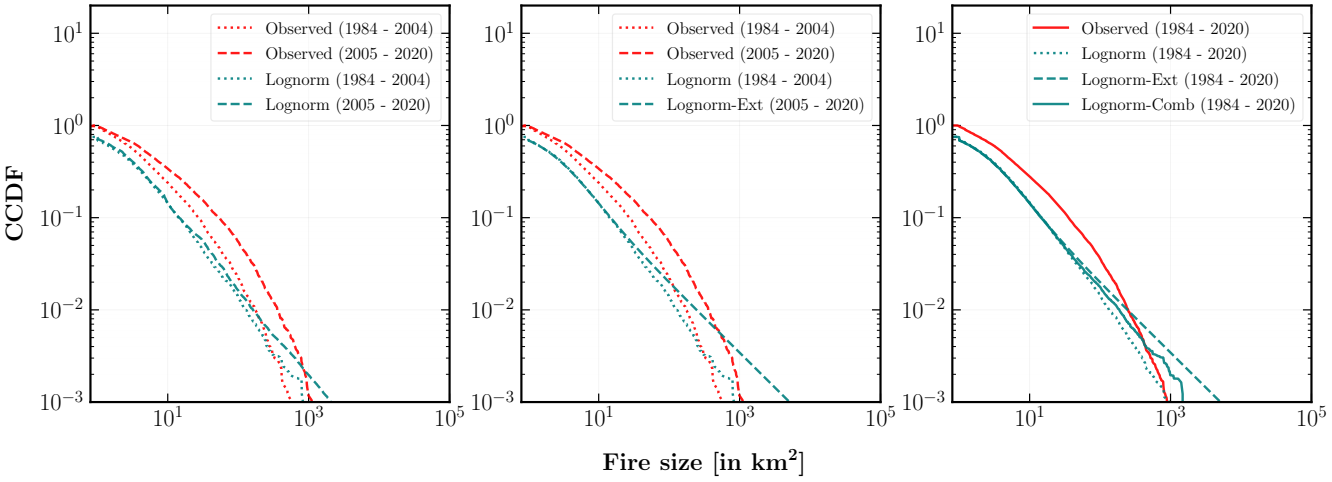

Figure S5. As in Fig. S4, but with a lognormal loss function for the MDN. Unlike the GPD, the lognormal distribution does not require a threshold on fire sizes.

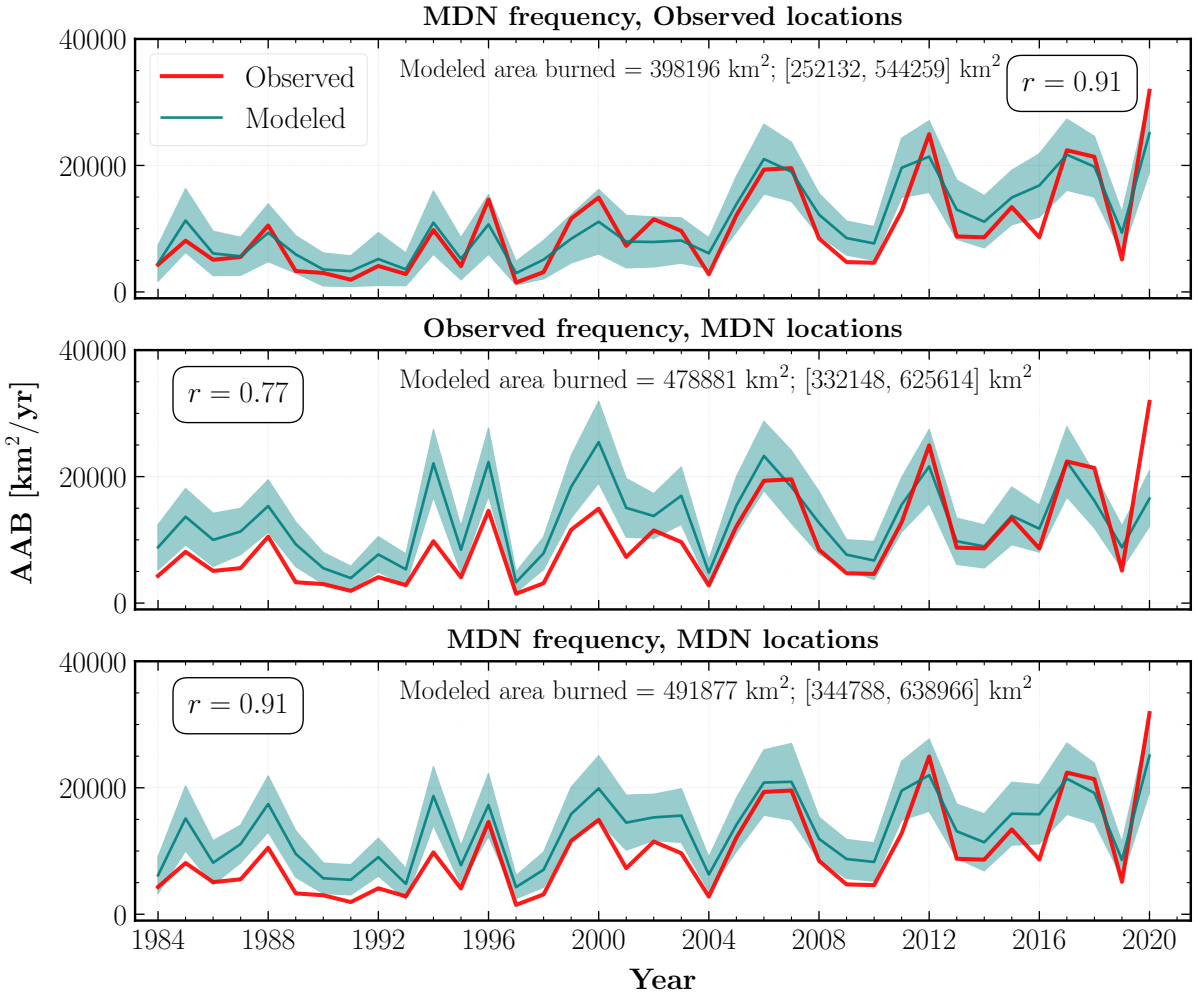

Figure S6. Cumulative observed (red) and modeled (teal) annual area burned (AAB) across the western United States from 1984 to 2020 for different combinations of fire frequencies and locations. The upper and lower panels show the AAB derived using modeled frequencies from the ZIPD MDN for each Ecoregion along with fire sizes simulated from the combined GPD model evaluated at observed and model fire locations respectively; 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.

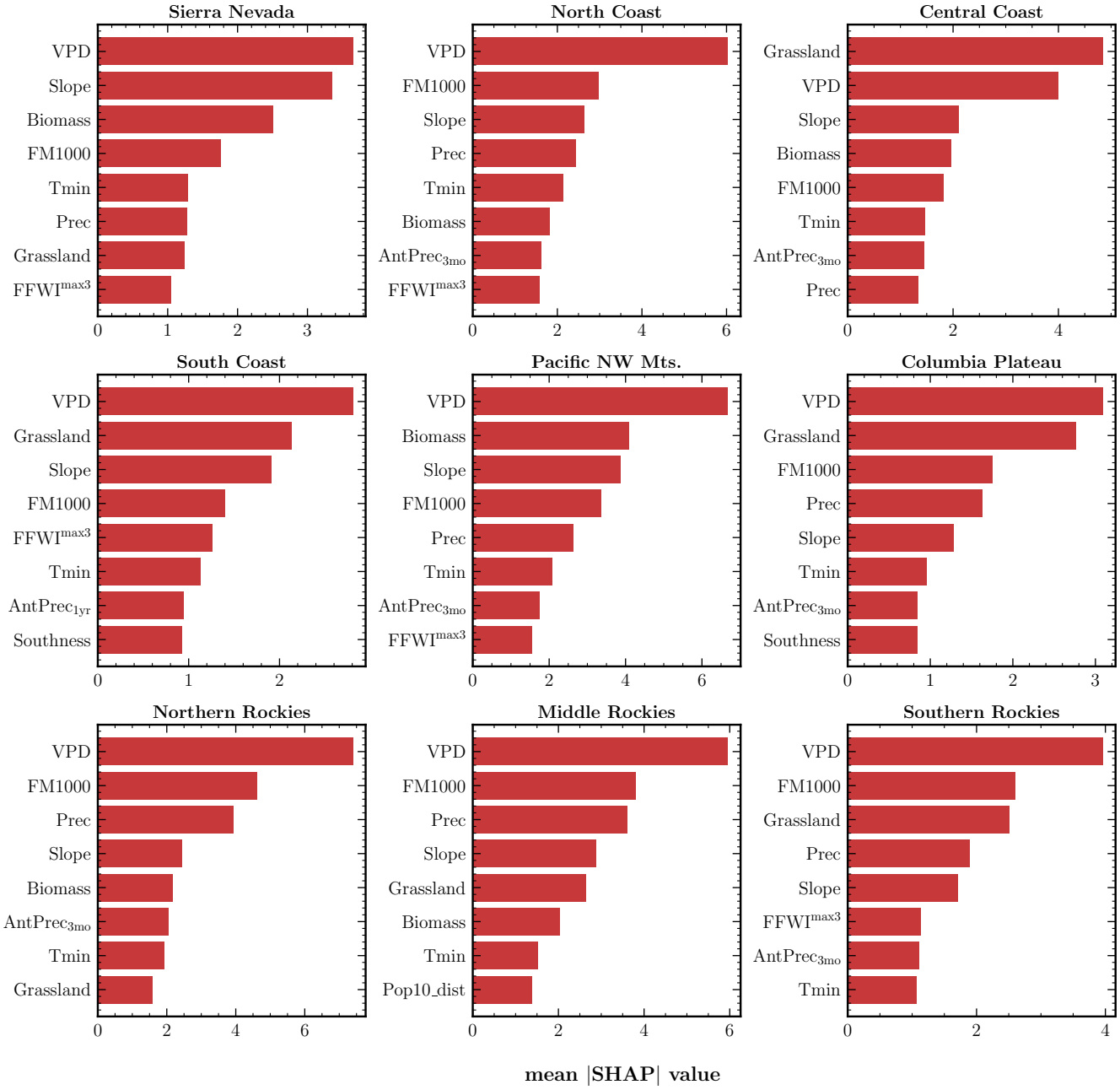

Figure S7. Mean SHAP values for the top 8 input predictors per ecoregion of the GPD size MDN. These include all the CA Ecoregions: Sierra Nevada, North, Central, and South Coasts; Pacific NW Mountains; Columbia Plateau; and North, Middle, and Southern Rockies.

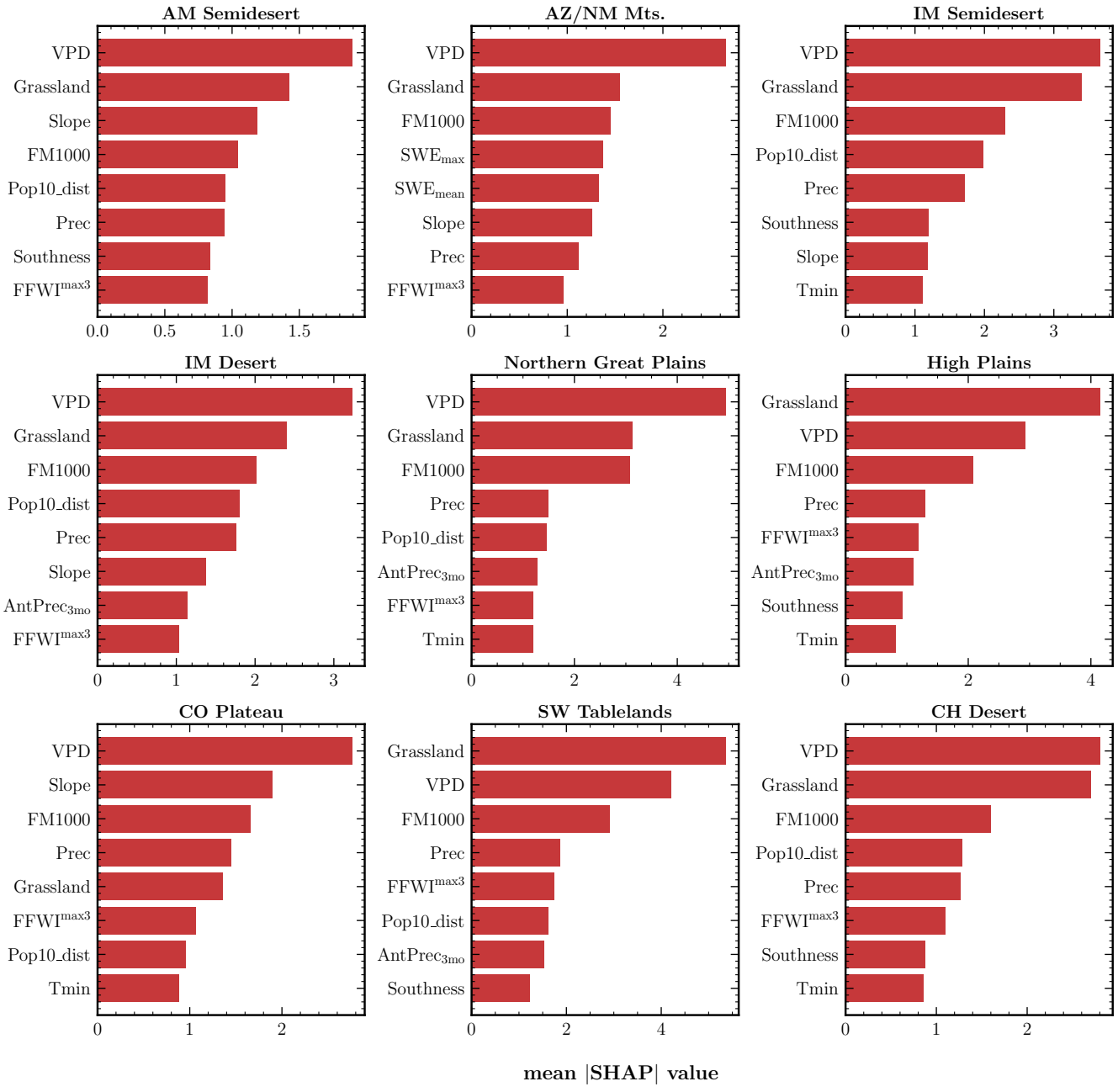

Figure S8. As in Fig. S7, but for the remaining WUS Ecoregions: American (AM) and Intermountain (IM) Semideserts, Arizona/New Mexico (AZ/NM) Mountains; Chihuahuan (CH) and IM Deserts; Northern Great and High Plains; Colorado (CO) Plateau; and Southwestern Tablelands.

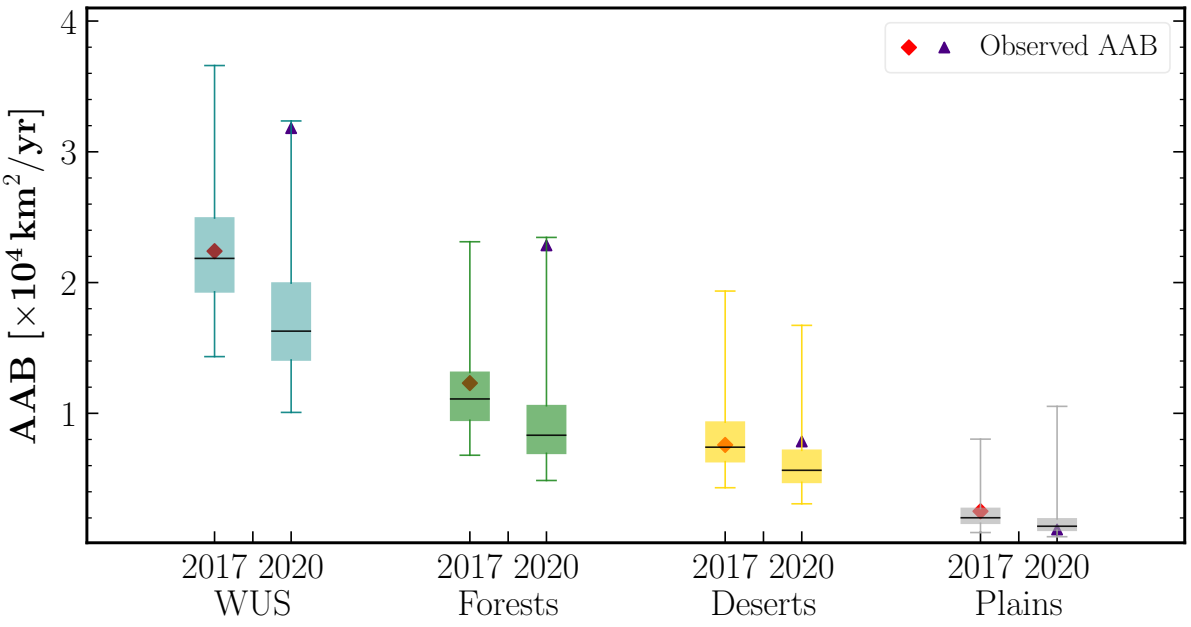

Figure S9. Boxplots of modeled annual area burned (AAB) for two extreme fire years, 2017 and 2020, for the entire western United States (WUS) (teal) and three Divisions organized by their primary vegetation types: Forests (green), Deserts (yellow), and Plains (gray). The lower and upper whiskers of each 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).

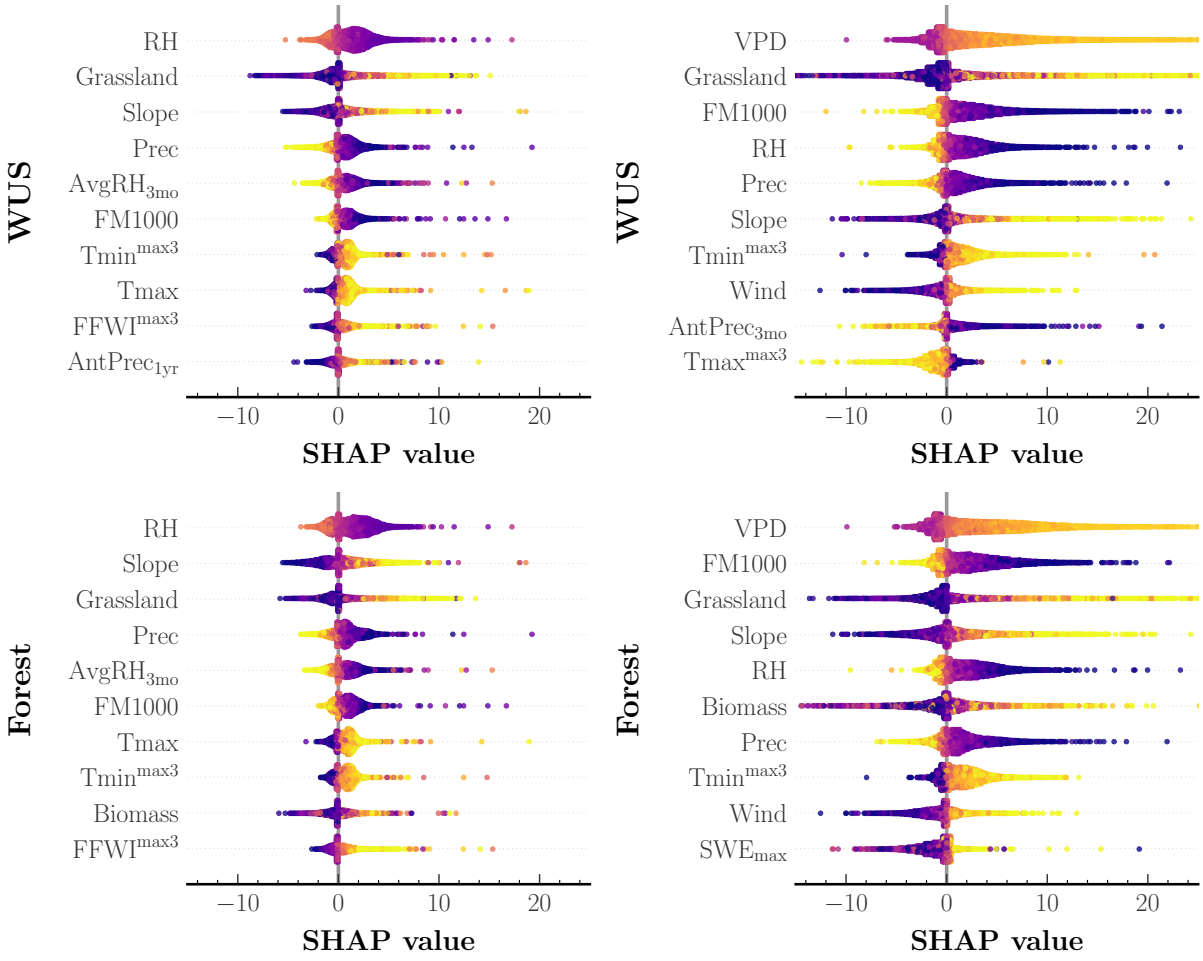

Figure S10. SHapley Additive exPlanation (SHAP) analysis of the fire size MDN model outputs for different sets of input predictors. *Left column:* SHAP summary plots with relative humidity (RH) and average RH over 3 antecedent months ($\mathrm{AvgRH}_{\mathrm{3mo}}$) predictors instead of their VPD counterparts 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.