







RESEARCH ARTICLE

Supplementary Material: Selecting Robust Features for Machine Learning Applications using Multidata Causal Discovery

Saranya Ganesh S.^{1*}, Tom Beucler¹, Frederick Iat-Hin Tam¹, Milton Gomez¹, Jakob Runge^{2,3} and Andreas Gerhardus²

¹Institute of Earth Surface Dynamics, University of Lausanne, Lausanne, Vaud, Switzerland

²Institute of Data Science, German Aerospace Center (DLR), Jena, Thuringia, Germany

³Faculty of Electrical Engineering and Computer Science, Technische Universität Berlin, Berlin, Germany

*Corresponding author. Email: saranyaganesh.s@gmail.com

Received xx xxx xxxx

In Section A, we included the full set of variables for our prediction problem. Section B of this supplemental information shows the results of the performance of causal models for the remaining targets, including Minimum Sea Level Pressure (MSLP) and Total Integrated Precipitation. Section B also includes results for experiments with a reduced lead time of 6 hours, using both the PC1 and PCMCI methods. The feature selection baselines for comparing the performances are defined in Section C, followed by the results for predicting the remaining targets. Finally, Section D shows the performance of the best models, as well as the causal predictors used in the model with the best skill on the validation set for maximum surface wind.

A. List of Variables Used as Predictors

We provide a list of all the variables chosen from the ERA5 (3 hourly) dataset, including targets and predictors, for preparing the ensemble of TC time series in Tab 1.

B. Optimal Number of Causal Features

Figures 3,4,5, & 6 show the comparison of M-PC₁ and M-PCMCI algorithms for the selected targets before and after temporally aligning the time series according to the time of minimum MSLP during the lifetime of each storms in the group. We see a clear improvement in the validation sets of the aligned dataset for both PC1 and PCMCI for all the targets.

C. Description of Machine Learning Algorithms

In this section, we provide a description of our implementations of the machine learning algorithms tested for this work. Prior to training the algorithms, we calculated the mean and standard deviation of each input feature in the available training data, noting that the values were considered for the set of all storms (and not on a per-storm basis). We then used these values to *standard-scale* the input features,

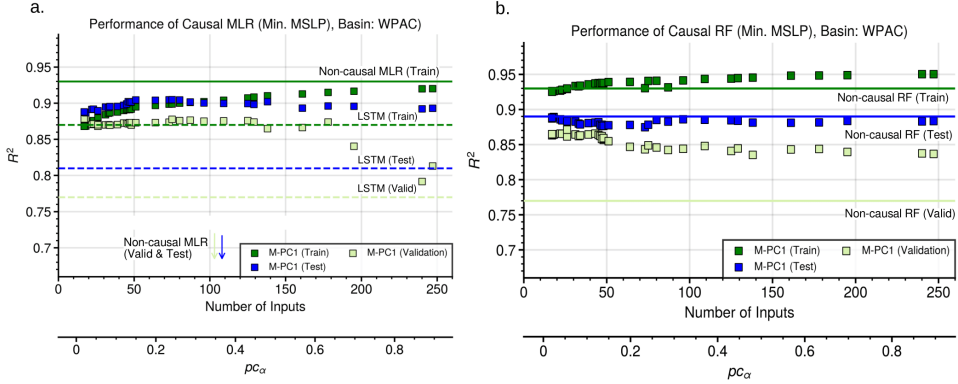


Figure 1. Performance of Causal ML for multiple tests by varying hyperparameter for the prediction of Minimum MSLP by causal-MLR (a) compared to noncausal ML (Solid lines) and LSTM (dashed lines), where as (b) shows the performance of Causal-RF compared to noncausal-RF (solid lines).

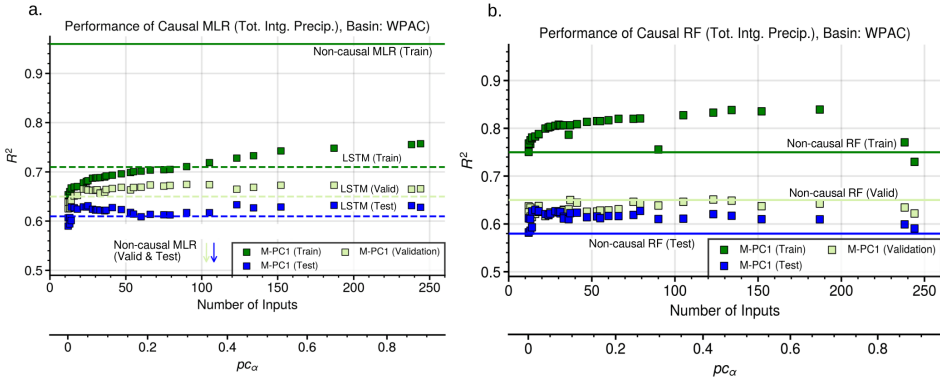


Figure 2. Same as Figure 1 but for the prediction of Total integrated Precipitation.

per Eqn. (1).

$$z_i = \frac{x_i - \mu_i}{\sigma_i} \quad (1)$$

Multiple linear Regression - In order to benchmark the performance of multidata causal feature selection, a plurality of multiple linear regression (MLR) algorithms were prepared, using the Scikit-Learn implementation of the Linear Regression algorithm and its corresponding default parameters. Each individual MLR algorithm was trained to predict one of three unscaled target variables (i.e., one of MSLP, precipitation, or Surface Wind) using the selected, standard-scaled inputs being evaluated.

Random Forest Regression - To ensure that the benefit of causal feature selection extends to more complex, nonlinear machine learning algorithms. We applied the same sets of input variables used to train the causal and non-causal MLR models to a Random Forest Regressor (RF Regressor). The implementation of the RF regression algorithm in this study utilizes that provided in the Scikit-Learn package. Compared to the MLR models, the RF Regressor contains several trainable hyperparameters that we can optimize for better prediction skills. Using the *RandomizedSearchCV* function, we tuned the hyperparameters related to the depth of the model, minimum number of samples to split decision trees, and the number of estimators. The best model that has the best cross-validation accuracy on the training data is chosen for analysis.

The noncausal feature selection baselines that are used in the main manuscript are described below.

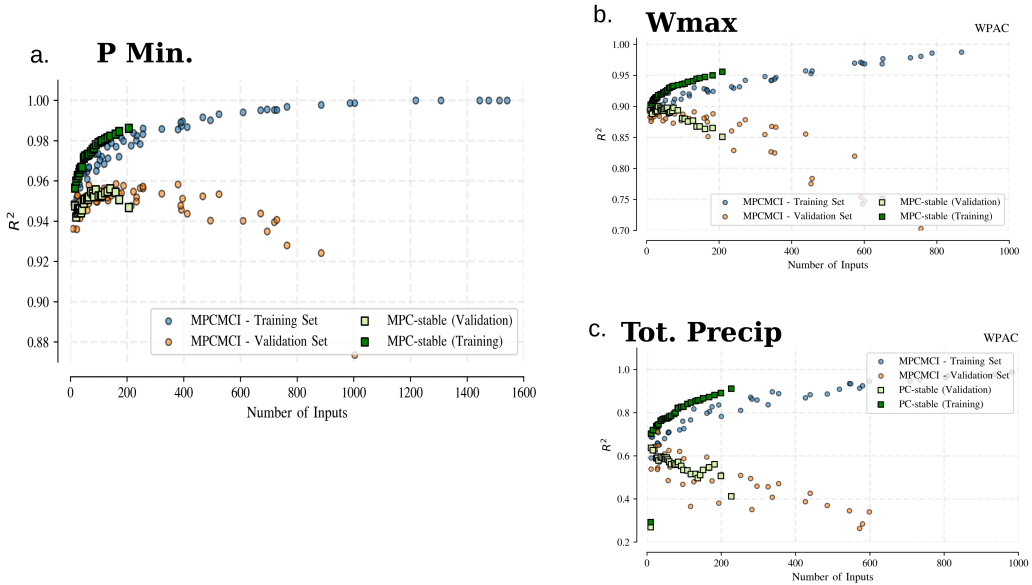


Figure 3. Comparison of $M-PC_1$ and $M-PCMCI$ based Causal-MLR model performance for 6 hour predictions without time alignment and with time lags ranging from 6 hrs to 2 days for selected targets.

Random Feature Selection is a sampling method where features are chosen randomly. Random sampling is analogous to drawing out a set of cards after shuffling without any criteria. Our implementation of this algorithm randomly selects a set of input features (size ranging from 10 to 1000) from all possible combinations of variables and time lags.

Lagged Correlation considers the absolute correlation between the prediction targets and different time-lagged input features. We adopted a kitchen sink approach where we obtained the correlation values between targets and all time-lagged variables by c . These correlation values are ranked and the features with the highest correlations are then chosen as MLR inputs. The size of these sets of features ranges from 10 to 1000.

XAI takes the training dataset to build a random forest regression model using Python’s scikit-learn library. By using this baseline method, we explore whether the use of feature importance (when nonlinear relationships between variables and targets are included) can result in a better selection of features. The Gini feature importance as measured by the trained random forest regressors provides an objective means to rank and select the most informative input variables. Input variables are ranked from most important to least important based on Gini impurity-based feature importances. The top-ranked features are then chosen to train the MLR models. Alternative feature importance methods, e.g., permutation feature importance or absolute Shapley values, are left for future work.

LSTM Neural Network - We prepared three Long Short-Term Memory (LSTM) recurrent neural networks as baselines, training the LSTM models on standard-scaled input data and configuring each LSTM to predict one of the standard-scaled target variables (i.e., one of MSLP, precipitation, or Surface Wind). We implemented each LSTM as a sequential model using PyTorch; their architecture includes an LSTM layer, a dropout layer, a linear hidden layer, and a linear output layer. As we targeted standard scaled outputs, the output of the network needed to represent positive and negative values. To do this, we set the output activation function to the identity function and we set the hidden layer activation function to hyperbolic tangent. We selected the Adam optimizer and mean-square error loss for our training, and proceeded to conduct a hyperparameter search using the Optuna framework. The study

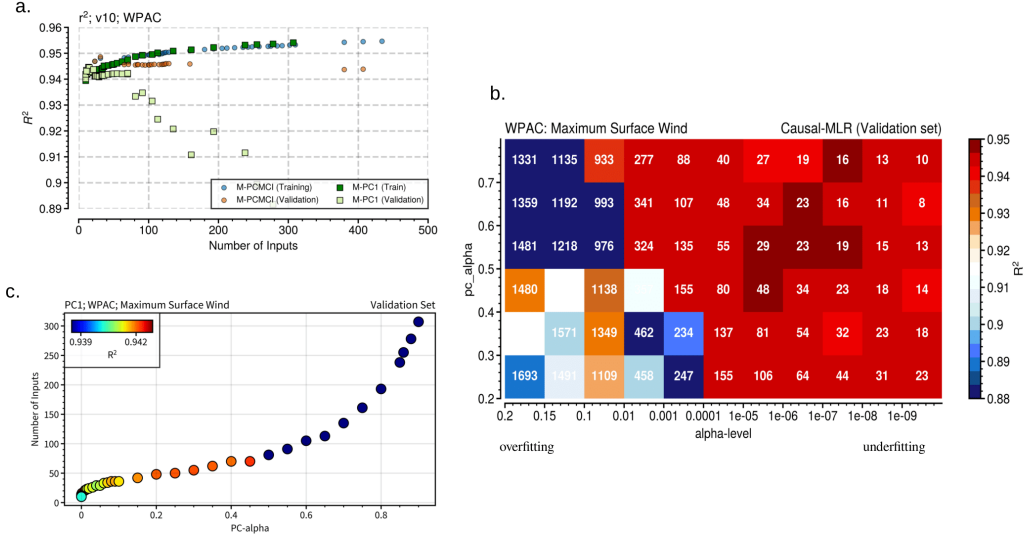


Figure 4. Performance of PC_1 and $PCMC1$ models with time aligned inputs for the prediction of Maximum Wind (a), the relationship between hyper-parameters, inputs, and performance (b, c).

employed 10 trials that tested LSTM and hidden layers 50-100 units wide, dropout rates between 0.0 and 0.5, and learning rates between $1e-4$ and $1e-3$. We note that we set the number of units in the LSTM layer and the hidden layer to be equal to each other in all conducted trials.

D. Comparison of Feature Selection Baselines

A comparison of the performance of Causal MLR to the performance of MLR models based on other feature selection baselines for the targets, Minimum MSLP and Total Integrated Precipitation are shown in Figures 7 and 8 respectively.

The Mean Square Error (MSE) and Mean Absolute Error (MAE) of the best model prediction of *Maximum Surface Winds*, *MSLP* and *Total integrated Precipitation* on both the training, validation and test sets for the best ML models. All metrics signify a good performance for Causal-ML with far less number of inputs compared to the number of inputs from the best models using the Non-causal-RF and Non-causal MLR methods. For the best ML models used, MSE are listed in Table 2 and MAE are listed Table 3.

E. Optimal Causal Predictors

The predictors and time lags for the best causal-MLR model with time-aligned inputs for the prediction of maximum wind speeds 1-day in advance are shown in Table 4.

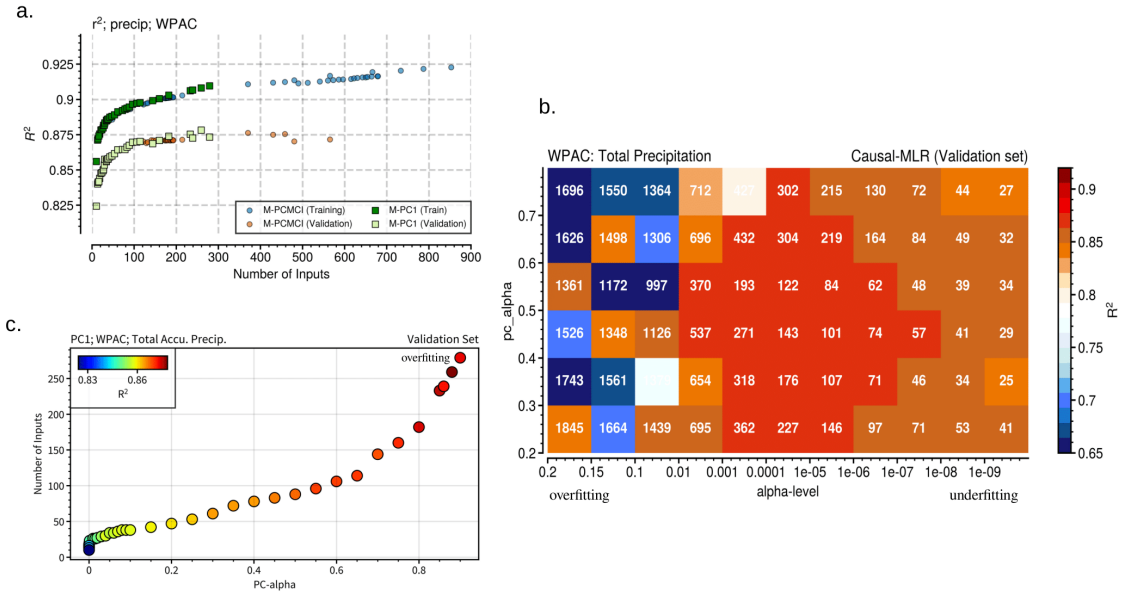


Figure 5. Same as the previous figure, but for the prediction of total integrated precipitation.

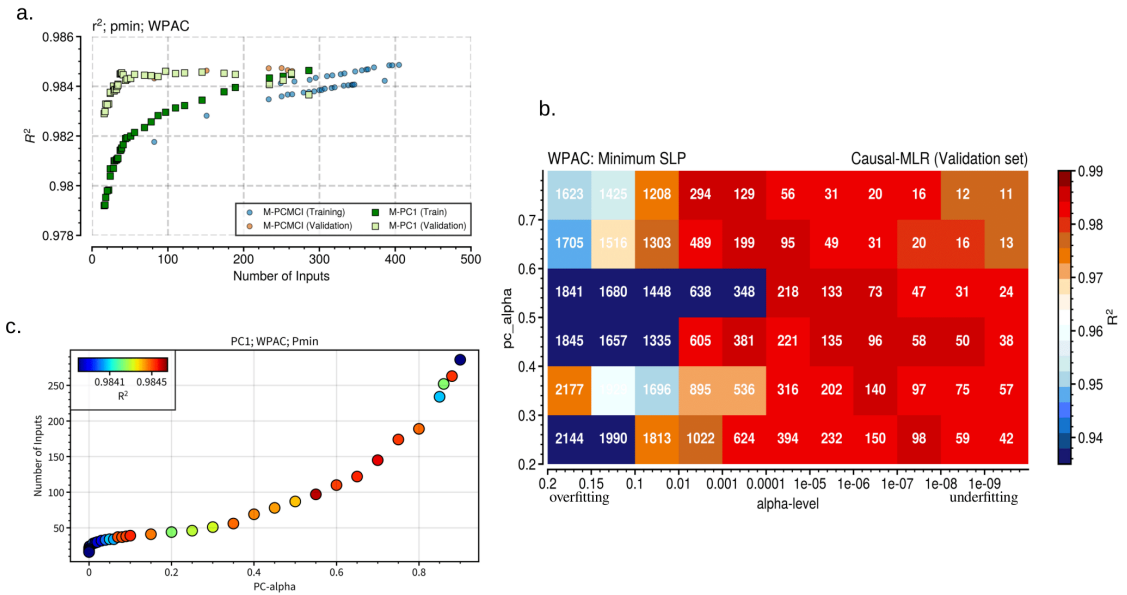


Figure 6. Same as the previous figure, but for the prediction of Minimum MSLP.

Targets (500 km radii)	Minimum Sea Level Pressure (Pmin)	Maximum Wind Speed at 10 m (V10)	Total Integrated Precipitation (Precip.)
Predictors : Inner core (200 km Radii) and Outer Core (200 to 800 km radii radii)	Vertical pressure levels(hPa)		Variable list
	1000, 975, 950, 925, 900, 850, 800, 700, 600, 500, 400, 300, 250, 200, 150, 100, 50		Relative Vorticity, Relative humidity, Geo-potential Height
	1000, 975, 925, 850, 700, 600, 500, 400, 300, 250, 200, 150, 100		Vertical Velocity
	1000, 925, 850, 800, 700, 600, 500, 400, 300, 250, 200, 100, 50		Horizontal Divergence
	1000, 925, 850, 700, 600, 500, 400, 300, 250, 200		Equivalent Potential Temperature
	Single levels		Dew Point Temperature – 2 m, Temperature – 2 m, Convective available potential energy, Sea Surface Temperature, Total column water vapor, Total column cloud ice water, Total column cloud liquid water, Total column super-cooled liquid water, Total column cloud rain water, Vertical integral of divergence of cloud frozen water flux, Vertical integral of divergence of cloud liquid water flux, Vertical integral of divergence of mass flux, Vertical integral of divergence of moisture flux, Vertical integral of divergence of total energy flux, Vertical integral of potential and internal energy, Vertical integral of potential internal and latent energy, Vertical integral of thermal energy, Vertical integral of total energy, Vertically integrated moisture divergence, Mean vertically integrated moisture divergence, Instantaneous moisture flux, Instantaneous surface sensible heat flux, Mean surface latent heat flux, Mean surface sensible heat flux, Surface latent heat flux, Surface sensible heat flux.
Predictors Outer Core (200 – 800 km radii)	Vertical Wind Shear		1000 hPa - 200 hPa, 1000 hPa – 300 hPa, 1000 hPa – 500 hPa, 1000 hPa – 700 hPa, 1000 hPa – 850 hPa, 850 hPa – 200 hPa, 850 hPa – 250 hPa, 850 hPa – 300 hPa, 850 hPa – 500 hPa, 925 hPa – 200 hPa, 925 hPa – 250 hPa.

Table 1. List of variables used from ERA5 dataset.

ML Models		Training (No. of features)			Validation			Test		
Target		Pmin (hPa)	V10 (ms ⁻¹)	Precip ×10 ⁻³ (km ²)	Pmin	V10	Precip	Pmin	V10	Precip
Causal RF		13.45 (26)	3.49 (17)	35.45 (123)	28.05	6.24	63.3	24.03	6.26	92.1
Causal MLR		24.67 (17)	5.2 (31)	61.38 (90)	25.3	5.62	59.01	21.89	5.87	92.79
Non-causal RF	All	15.68 (3978)	3.91 (3978)	46.25 (3978)	34.92	7.58	100.03	24.39	6.54	102.16
	Lagged	8.80 (480)	2.3 (560)	37.8 (80)	29.22	6.23	82.12	21.04	5.65	93.2
	Random	8.57 (870)	2.27 (770)	27.94 (970)	38.51	7.8	105.23	27.15	7.19	105.77
Non-causal MLR	All	2.43 (3978)	0.66 (3978)	7.87 (3978)	373.54	398.33	33370.94	105.74	31.03	338.29
	Lagged	15.64 (440)	12.04 (40)	60.60 (120)	31.09	14.77	91.40	17.20	11.83	85.70
	Random	19.05 (420)	5.58 (130)	58.23 (290)	41.87	8.20	97.68	29.88	7.81	110.5
	XAI	16.93 (240)	3.84(420)	55.34 (140)	30.90	6.05	80.32	19.92	6.38	90.7
LSTM		27.55 (3978)	6.49 (3978)	179.76 (3978)	44.00	8.80	199.29	39.44	8.02	206.12

Table 2. MSE.

ML Models		Training (No. of features)			Validation			Test		
Target		Pmin (hPa)	V10 (ms ⁻¹)	Precip ×10 ⁻³ (km ²)	Pmin	V10	Precip	Pmin	V10	Precip
Causal RF		2.55 (26)	1.42 (17)	0.14 (123)	3.87	1.94	0.19	3.62	1.97	0.23
Causal MLR		3.49(17)	1.77 (31)	0.19 (90)	3.62	1.84	0.19	3.49	1.89	0.23
Non-causal RF	All	2.81 (3978)	1.50 (3978)	0.16 (3978)	4.11	2.17	0.23	3.65	1.99	0.24
	Lagged	2.04 (480)	1.12 (560)	0.14 (80)	3.78	1.95	0.22	3.41	1.85	0.23
	Random	2.07 (870)	1.12 (770)	0.12 (970)	4.43	2.2	0.25	3.94	2.1	0.25
Non-causal MLR	All	1.21 (3978)	0.63 (3978)	0.07 (3978)	10.42	6.15	0.85	7.71	4.44	0.44
	Lagged	2.90 (440)	1.54 (40)	0.17 (120)	3.74	2.07	0.22	3.11	1.70	0.21
	Random	3.37 (420)	1.56 (130)	0.19 (290)	5.11	2.32	0.24	4.14	2.17	0.24
	XAI	3.03 (240)	1.76 (420)	0.17 (140)	4.07	1.97	0.22	3.4	1.93	0.22
LSTM		3.90 (3978)	1.97 (3978)	0.34 (3978)	4.85	2.29	0.35	4.74	2.22	0.36

Table 3. MAE (lower values means better models)

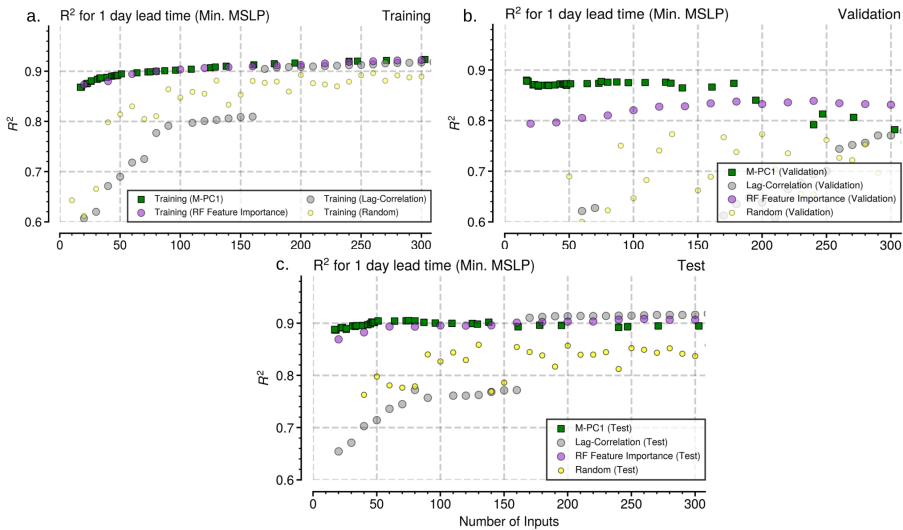


Figure 7. Comparison of the performance of Training, Validation and Test sets of MLR models that used different feature selection methods for predicting minimum MSLP.

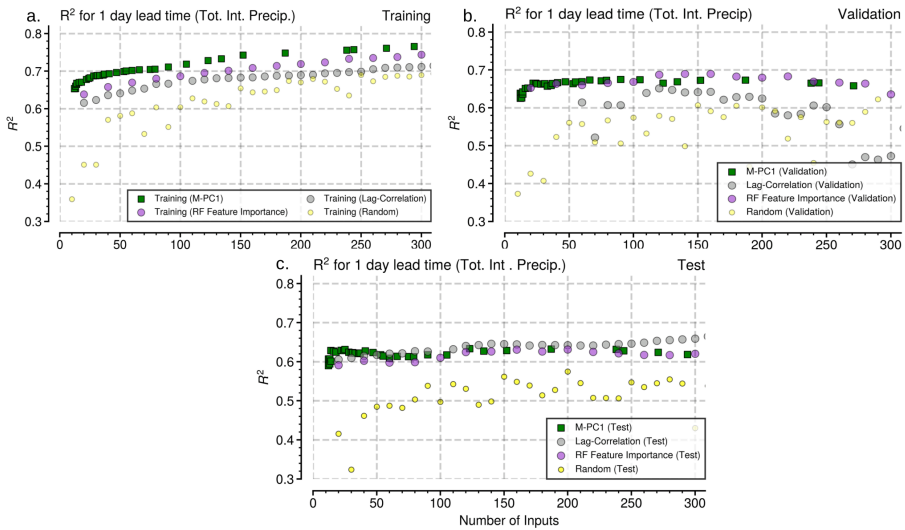


Figure 8. Same as Figure 7. but for predicting Total area integrated precipitation.

Causally linked predictors from best ML model for Maximum Wind Speed (10 m)					
Time lags in hours (3 hrly interval)	Single level		Multiple level		Others
	Inner	Outer	Inner	Outer	
24	<ol style="list-style-type: none"> 1. Convective Available Potential Energy 2. Total column cloud ice water 3. Instantaneous Sensible heat flux 4. Vertically integrated divergence of mass flux 5. Total column water vapor 6. Instantaneous moisture flux 	<ol style="list-style-type: none"> 1. Vertical integral of thermal energy, 2. Vertical integral of potential internal and latent energy 3. Vertical integral of total energy 	1. Divergence – 850 hPa	<ol style="list-style-type: none"> 1. Divergence – 925 hPa 2. Relative humidity – 950 hPa 	<ol style="list-style-type: none"> 1. Minimum SLP 2. Maximum Wind Speed
27			<ol style="list-style-type: none"> 1. Divergence – 850 hPa, 2. Relative humidity –1000 hPa 	1. Relative Vorticity – 700 hPa	1. Minimum SLP
30		1. Instantaneous moisture flux			
33			<ol style="list-style-type: none"> 1. Relative humidity – 500 hPa 2. Relative humidity – 100 hPa 	1. Geopotential height – 800 hPa	
36	1. Convective Available Potential Energy		1. Vertical velocity – 1000 hPa		
39			1. Vertical velocity – 1000 hPa	1. Equivalent Potential Temperature - 400 hPa	1. Vertical shear 1000 – 700 hPa
42				1. Divergence –300 hPa	1. Vertical shear 1000 - 850 hPa
57				1. Divergence –200 hPa	
66			1. Relative humidity –1000 hPa		

Table 4. List of 31 causally linked predictors for Maximum wind (1 day lead-time) at significant time lags with best model using PC₁