MULTIDATA CAUSAL DISCOVERY FOR STATISTICAL HURRICANE INTENSITY FORECASTING

Saranya Ganesh S., Frederick Iat-Hin Tam, Milton S. Gomez, Tom Beucler
Faculty of Geosciences and Environment,
Expertise Center for Climate Extremes,
University of Lausanne
Lausanne, Vaud, Switzerland

Marie McGraw, Mark DeMaria, Kate Musgrave
Cooperative Institute for Research in the Atmosphere,
Colorado State University,
Fort Collins, Colorado

Jakob Runge
Department of Computer Science,,
University of Potsdam
Potsdam, Germany

ABSTRACT

Improving statistical forecasts of Atlantic hurricane intensity is limited by complex nonlinear interactions and difficulty in identifying relevant predictors. Conventional methods prioritize correlation or fit, often overlooking confounding variables and limiting generalizability to unseen tropical storms. To address this, we leverage a multidata causal discovery framework with a replicated dataset based on Statistical Hurricane Intensity Prediction Scheme (SHIPS) using ERA5 meteorological reanalysis. We conduct multiple experiments to identify and select predictors causally linked to hurricane intensity changes. We train multiple linear regression models to compare causal feature selection with no selection, correlation, and random forest feature importance across five forecast lead times from 1 to 5 days (24–120 hours). Causal feature selection consistently outperforms on unseen test cases, especially for lead times shorter than 3 days. The causal features primarily include vertical shear, mid-tropospheric potential vorticity and surface moisture conditions, which are physically significant yet often underutilized in hurricane intensity predictions. Further, we build an extended predictor set (SHIPS+) by adding selected features to the standard SHIPS predictors. SHIPS+ yields increased short-term predictive skill at lead times of 24, 48, and 72 hours. Adding nonlinearity using multilayer perceptron further extends skill to longer lead times, despite our framework being purely regional and not requiring global forecast data. Operational SHIPS tests confirm that three of the six added causally discovered predictors improve forecasts, with the largest gains at longer lead times. Our results demonstrate that causal discovery improves hurricane intensity prediction and pave the way toward more empirical forecasts.

Keywords Causal Feature Selection · Machine Learning · Statistical Hurricane Intensity Prediction · Tropical cyclones

1 Introduction

Extensive research on tropical cyclones (TC), including cyclogenesis and intensification, has advanced our understanding of key atmospheric and oceanic processes [1, 2, 3]. However the continued rise in coastal populations, and the increasing risks of wind gusts, storm surge, extreme rainfall, and severe weather [4, 5] underscore the need for more accurate and robust predictions of TC intensity.

Despite breakthroughs in short-range to subseasonal prediction regimes [6, 7, 8], accurately predicting rapid storm intensity changes remains challenging [9, 10, 11] especially after 24 hours, due to error growth in initial conditions, inherent misrepresentations of the model due to lack of understanding of the underlying processes, and inadequate data assimilation techniques [12]. In addition to numerical weather prediction systems, statistical predictive models play a significant role in TC forecasts by contributing to the improved predictive skill of multi-model ensemble-based consensus forecasts [13]. However, both statistical and dynamic models struggle to estimate the rapid intensification of TCs, which is a major challenge for storm forecasters, as multiscale air-sea interactions and radiative-convective feedback leading to rapid changes in intensity are still not fully understood [14]. However, when comparing probabilistic predictions from high-resolution models and statistical schemes, the latter outperform dynamical models in predicting rapid intensification [15]. The forecast skill of the Statistical Hurricane Intensity Prediction Scheme (SHIPS), which integrates large-scale predictors from climatology, persistence, and synoptic predictors to estimate hurricane intensities[16, 17] has gradually improved over the years. SHIPS evolved from a "statistical-synoptic" to a "statistical-dynamical" model when it started using synoptic environmental conditions from dynamical models (e.g. GFS) in statistical algorithms [13]. The combined consensus forecast has the best skill for TC intensity prediction, therefore SHIPS adds value to operational intensity prediction efforts [18]. Although SHIPS includes predictors from ocean analyses and satellite imagery, the forecast skill diminishes with lead time, partly because predictors are chosen semiempirically based on domain knowledge, climatology and persistence [19]. To increase sample size, the regression coefficients for SHIPS are re-derived after each hurricane season since 1993 [11], including annual climatological adjustments and continual updates to the predictor list. However, while the SHIPS developmental dataset provides an extensive set of environmental predictors, it may still omit key variables relevant for changes in TC intensity.

In this study, we address this challenge by adopting a causal discovery framework [20, 21, 22, 23, 24] to objectively identify and suggest new environmental predictors to be added to the latest versions of the SHIPS developmental dataset. Recent work in climate science used linear and nonlinear causal discovery methods [25, 26, 22] to unveil spurious associations in existing statistical prediction models due to common drivers or indirect associations. Once the causal graph is known, the strength of the links between the physical variables can be determined with causal inference [27], complementing data-driven techniques [24]. In this study, we used the statistically significant causal links identified with causal discovery frameworks to determine predictors with direct causal links to changes in TC intensity.

Despite this recent progress, causality is still commonly inferred via lagged correlation analysis. The main disadvantage of lagged correlation analysis is that it does not provide insight into the causal directions of the relationships between variables [28]. As such, lagged correlation methods are not immune to non-causal correlations from autocorrelation effects, indirect connections through a third process, or a shared driver. Such non-causal and spurious correlations hinder interpretability in statistical prediction models [29, 30]. Understanding the directional causality between atmospheric variables requires additional methodologies or causal inference approaches beyond the scope of lagged regression analyses. [31] use the graphical model-based structure learning approach of the PC causal discovery algorithm to identify key variables for tropical cyclogenesis prediction. They found that predictive modeling using logistic regression employing the highest ranked variables improved statistical predictive skills; their results further suggest that causal methodologies could be used to create new tropical cyclogenesis indices. Additionally, the PCMCI+ causal algorithm — a combination of the PC algorithm and the Momentary Conditional test [32] — have been used to identify precursor regions useful for improving seasonal hurricane frequency forecasts [33].

Here, we leverage the PC algorithm and increase statistical power by applying the recently developed *multidata* causal discovery framework to suggest new predictors for the operational SHIPS model. This approach assumes that all Atlantic hurricane intensity changes are governed by a shared causal graph, which is reasonable given that different hurricanes represent multiple realizations of the same physical phenomenon. The multidata PC algorithm requires multivariate time series of all potential predictors for each storm, motivating our replication of the SHIPS developmental dataset using meteorological analyses, as described in Section 2. After detailing our causal methodology and experimental design in Section 3, we demonstrate in Sections 44.1 and 44.2 that causal feature selection outperforms correlation-based methods in identifying useful predictors for TC intensity forecasts, particularly at short lead times. These improvements are robust in both reanalysis-based experiments (Section 44.3), where we also show that the causal predictors nonlinearly drive intensity (Section 44.4), and in operational-like settings using real-time global model analysis and forecast fields (Section 44.6). In Section 44.5, we present a case study of Hurricane Larry to illustrate the physical relevance of the causal links between the selected predictors and short-term TC intensification.

2 Data Sources and Preprocessing

Our analysis is based on two complementary experimental designs: (i) a hierarchy of reanalysis-based experiments using the high-resolution ECMWF Reanalysis Version 5 (ERA5) [34, 35], and (ii) a developmental SHIPS dataset experiment that tests the added value of these causally informed predictors alongside the operational SHIPS predictors. Our causal discovery experiments focus solely on SHIPS predictors that can be replicated using ERA5 variables; satellite-based predictors such as GOES brightness temperature and Ocean Heat Content (OHC), which are not directly reproducible in ERA5, are excluded to ensure a fair comparison and reliable interpretation of causal relationships based on ERA5 data. To define TC-centered domains and intensity change targets, we rely on the International Best Track Archive for Climate Stewardship (IBTrACS), which provides reference track and intensity data for each TC.

2.1 IBTrACS: Reference Dataset for Intensity and Track

IBTrACS [36] provides a comprehensive record of TC tracks and intensity estimates compiled from all Regional Specialized Meteorological Centers and other TC warning agencies, including the Joint Typhoon Warning Center. We use IBTrACS in two complementary ways. First, the 6-hourly best-track locations define the TC-following domains used to extract area-averaged ERA5 synoptic predictors. Second, the observed maximum sustained surface (10 m) wind speeds are compared in each forecast lead time (24, 48, 72, 96 and 120 hours) to create the DELV24, DELV48, DELV72, DELV96, and DELV120 intensity change targets (in units m/s). TC intensity is commonly estimated with the satellite-based Dvorak method, which is systematically biased for weak and intense hurricanes [37]. To minimize potential biases related to Dvorak intensity estimates, we restrict our analysis to the North Atlantic, where routine aircraft reconnaissance missions provide regular in situ measurements of the true TC intensity. We select 247 long-lived Atlantic hurricanes (2000–2021) with lifetimes of at least four days before landfall for analysis.

2.2 ERA5 and TC PRIMED Datasets

The ERA5 predictor input to the causal feature selection algorithm consists of 6-hourly, area-averaged time series extracted from TC-following domains centered on IBTrACS best-track positions. To capture both inner-core storm dynamics and broader environmental conditions, predictors are computed for multiple radial regions around the storm center. Specifically, inner-core variables are averaged over 0–2° (approximately 0–200 km), while outer-area predictors are averaged from 200–800 km or up to 1000 km, following conventions used in the extended SHIPS developmental predictors. To supplement these ERA5-derived predictors, we also make use of the Tropical Cyclone Precipitation, Infrared, Microwave, and Environmental Dataset (TC PRIMED) [38]. TC PRIMED uses ERA5 to reconstruct operational SHIPS predictors as well as environmental and thermodynamic variables across multiple pressure levels and radial regions. This dataset provides an expanded pool of predictors beyond those we directly derived from ERA5. For reference, see the documentation and the products.

As the first set, we have the original operational predictors from the SHIPS developmental dataset. For causal experiments, we replicate these core SHIPS developmental predictors, which are derived directly from ERA5 reanalysis fields (e.g., VMAX, POT, PER, vertical wind shear; see Table S2 for details), providing a direct comparison to operational baselines. This is complemented by a broader pool of synoptic and thermodynamic variables sampled at multiple standard pressure levels and area-averaged within the defined radial domains, using ERA5 and supplemented by variables from TC PRIMED. These additional variables include divergence, vorticity, potential vorticity, equivalent potential temperature, geopotential height, relative humidity, air temperature, temperature gradients, precipitable water, and warm-core anomalies, among others. A complete list of variable names, radial averaging details, and pressure levels is provided in the Supplementary Information: Table S1 (original SHIPS predictors), Table S2 (ERA5 replication), Table S3 (inner-core variables), Table S4 (outer-area variables), and Table S5 (variables from TC PRIMED).

A key requirement for applying causal discovery to time series is the assumption of *causal stationarity*, i.e., the causal relationships between variables remain invariant over time within a given TC. This allows pooling of statistical evidence across time using a sliding-window approach for conditional independence testing. Here, we also assume that the causal relationships governing TC intensity change are consistent across storms, enabling conditional independence tests to be applied across the full training set [39]. To increase the likelihood of satisfying this assumption, we align TC time series according to their life cycle. Specifically, we aligned the Mean sea level Pressure (MSLP) time series for each storm in the training set relative to the time of minimum central pressure, smoothed using a Gaussian filter ($\sigma = 3 \times 6$hr) to reduce noise. The idea here is to align these unique indices together to satisfy causal stationarity. For the shorter time series, we append nan values on either sides to make sure that the lengths are consistent with the longest time series with minimum pressure in the middle. We aligned the time series using these indices so that the evolution of different storms could be compared on a common reference frame. All predictors are then standardized using the mean and standard deviation computed over the training set. The final ERA5 dataset includes 214 predictors and the

intensity-change target (DELV) at each forecast lead time. TCs from 2000–2019 are split into non-overlapping training and validation sets (85% and 15%, respectively), while an independent test set includes TCs from 2020–2021, along with Hurricane Wilma (2005), reserved as a particularly intense and challenging case.

2.3 SHIPS Developmental Dataset

To evaluate how causally informed predictors improve operational forecasting, we extract SHIPS predictors for the same 247 North Atlantic TC cases (2000-2021) used in the ERA5 experiments. Importantly, we only used predictor values from the 00-hour initialization time in the SHIPS developmental dataset. These values are derived from GFS analysis fields, which represent the initial-time atmospheric conditions (i.e., observations assimilated into the GFS model) at forecast hour zero. Unlike GFS forecast fields that estimate future atmospheric states, analysis fields reflect the best available estimate of the atmosphere at the initialization time. This setup places us in a purely statistical—rather than statistical-dynamical—framework, enabling a direct analogy between the SHIPS developmental dataset and the ERA5-derived dataset used for causal predictor discovery. We compare the generalization skill of statistical models trained on the original 21 SHIPS predictors ("original SHIPS") with models trained on the same predictors augmented with causally selected predictors ("SHIPS+"). These additional predictors were identified using ERA5-based causal discovery experiments but are computed from GFS analysis fields in exactly the same way as the original SHIPS predictors. This setup tests whether the predictors selected in the ERA5 replication setting can improve forecast skill when integrated into the operational SHIPS framework.

3 Methodology

3.1 Causal Feature Selection

We methodologically innovate by adopting the open-source TiGraMITe causal discovery package [20, 22] to find potential predictors of hurricane intensity change at different lead times. TiGraMITe has been widely used to discover causal relationships in climate data [40, 41, 42]. The multidata PC (M-PC) algorithm, a modified version of the first step of the Peter Spirtes and Clark Glymour Momentary Conditional Independence (PCMCI+) algorithm [22, 32], is used to discover time-lagged causal relationships in time-series data. Compared to the more advanced PCMCI+, which aims to learn the full causal graph including contemporaneous causal links, this work uses only the PC algorithm to eliminate potential spurious associations in lagged statistical prediction models [43, 44]. The "M" in M-PC indicates the multidata approach, where time series from multiple hurricanes are analyzed together to identify a consistent set of causal predictors. Clear non-causal drivers within the 215-variable ERA5 predictor dataset are detected and excluded through conditional independence tests, resulting in a single set of predictors that remain relevant across all training TCs [39]. In our setup, this is done without considering time-lagged variables — each predictor is evaluated at a single time-step for the given forecast lead time. Here we approximate conditional independence relations through a linear partial correlation test. By filtering out non-causal predictors in the statistical models, we aim to improve the generalizability of such models for unseen TC cases.

The M-PC algorithm's hyperparameters include the minimum and maximum time lags (τ_{min}, τ_{max}) and the statistical significance threshold for the partial correlation conditional independence tests pc_α to decide on variable removal. We set the minimum and maximum time lag in this study to $\tau_{min} = 4$ steps (1 day) and $\tau_{max} = 20$ steps (5 days) to discover potential novel TC predictors. The M-PC algorithm ranks the strength of the causal relationship between different predictors and TC intensity change by the absolute partial correlation value. Predictors are considered non-causal if the partial correlation between the variable and TC intensity change, conditional on the remaining variables, is non-significant. The names of causal predictors and their relevant time lags are stored in Python dictionaries.

3.2 Validation procedure

As causal discovery algorithms are sensitive to hyperparameter choice [45], the output of the linear causal feature selection framework should not be taken as truth for a complex physical problem such as TC intensity, but rather as a plausible guess of the true causal relationships. Careful cross-validation of the causal predictor lists is necessary to establish their robustness and trustworthiness. We postulate that causal predictors can be validated by the performance of the trained models on the validation data unseen during training. Our working assumption, supported by previous work [39], is that model generalization skill improves when adding causally relevant predictors, and degrades as spurious or redundant features are added. As such, our framework should identify a unique feature set that maximizes validation skill and highest likelihood of being causally related to TC intensification. We use 7-fold cross-validation with 31 storms per split, resulting in $31 \times 6 = 186$ storms for training and 31 storms for validation per fold. This yields multiple candidates of plausible causal relationships in the dataset. For one cross-validation fold, we train a series of

regression models based on PC output (causal variable names and relevant time lags) at different statistical significance thresholds (pc_α). Smaller pc_α values typically yield fewer predictors, though not always, which can result in different sets with the same cardinality. We assess the generalization capabilities of the trained MLR models with the coefficient of determination (R^2) and Pearson's Correlation Coefficient (PCC).
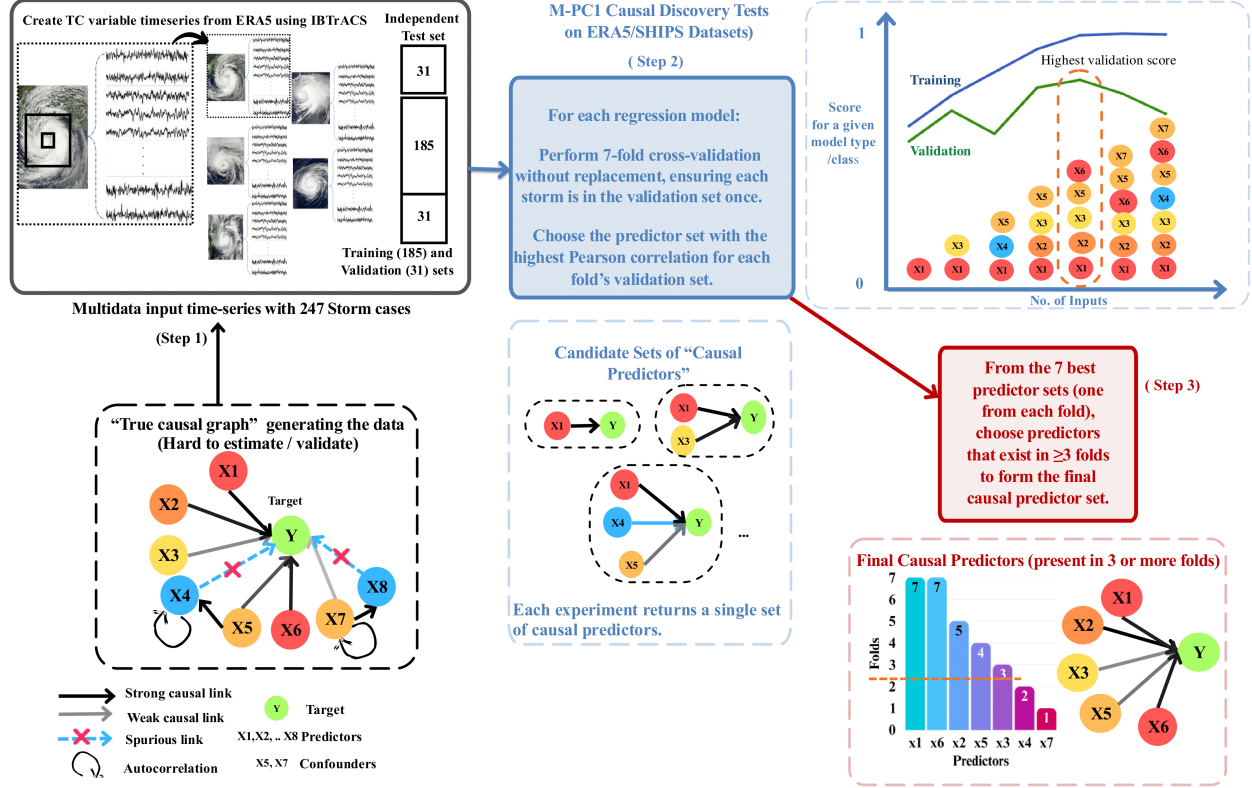

Figure 1: Multidata causal feature selection methodology. Step 1: Preprocessed spatiotemporal fields for all TC cases form an ensemble of *aligned* time series, which may contain spurious or non-causal relationships due to autocorrelation or confounding. Step 2: These multivariate time series (training set) are input to the multidata causal discovery algorithm (M-PC), which selects candidate predictors while controlling set size via hyperparameters. Each candidate set is evaluated using cross-validated regression. Step 3: Predictors appearing in at four out of the seven folds are pooled to form the final feature set. The goal is to estimate the portion of the true causal graph that helps predict TC intensity changes.

3.3 Filtering new potential predictors

Figure 1 shows our framework to discover new predictors using causal discovery algorithms. For all MLR models trained, we calculate the coefficient of determination R^2 for training, validation, and test sets. For each fold, we choose the model with the highest validation performance (R^2 and PCC), resulting in a set of 7 best models for analysis (Upper-right panel in Figure 1). Summarizing the R^2 of all trained models allows us to identify the most generalizable MLR model as the MLR model with the best R^2 score on the validation TCs. Adding more predictors beyond the set of inputs in the most generalizable MLR model will lead to model overfitting on the training TCs. These variables make the MLR model less generalizable because they constitute spurious associations arising from autocorrelation, indirect effects, or common drivers. To ensure robustness across cases, the final causal variable list presented in subsequent sections is produced by aggregating the variables used by the 7 best models. Only predictors that appear in more than 3 out of 7 best models and are not part of the existing SHIPS developmental predictor list will be shortlisted as new candidate predictors (lower right panel in Figure 1).

3.4 Experimental Design

Different experiments are conducted to demonstrate the potential applications of causal feature selection algorithms in complementing or enhancing existing statistical models for TC intensity change.

3.4.1 ERA5-based experiments for predictor discovery

The first set of experiments involves discovering candidate causal predictors from ERA5 reanalysis data that can be incorporated into the operational SHIPS. We designed two experiments to see whether causal feature selection framework can discover such variables and the sensitivity of discovered variables to the PC algorithm's assumptions. The first experiment (withASSUMPS) assumes that we will not remove any existing SHIPS predictors but only add new ones, representing a least-change case that is more feasible operationally. The second experiment (noASSUMPS) is a free-run, "kitchen-sink" scenario where the algorithm is allowed to overlook existing SHIPS predictors if they do not have strong causal relationships to TC intensity. We switch between the two scenarios in our framework by changing the predefined causal links in the dictionary input fed to a parameter (link_assumptions) in the Tigramite package. The link_assumptions input in the withASSUMPS experiments is edited to have *preexisting* causal links between the operational SHIPS predictors and the rate of TC intensity change at all times, regardless of model settings. In contrast, the noASSUMPS experiments do not contain predefined causal links and perform M-PC calculation fairly on all available predictors. Through this experiment, we investigate improving the generalizability of TC statistical models by removing spurious associations in the original SHIPS predictors.

3.4.2 Validation in an operation-like setting

The second set of experiments assesses the practical value of the causal predictors identified from the ERA5 reanalysis by testing them within the operational SHIPS framework. Specifically, an additional set of six diagnostic variables is extracted from the GFS fields, guided by the most robust predictors shortlisted from the ERA5 discovery experiments. To validate whether these new predictors improve generalization skill in an operational context, we replicate the same experimental setup described previously — comparing models trained with the original SHIPS predictors against models that incorporate the newly suggested predictors. This comparison directly quantifies whether enriching the SHIPS developmental dataset with causally selected predictors yields measurable improvements in forecast performance on unseen TC cases.

3.5 Regression Model hierarchy

3.5.1 Mapping and Optimization Objective

Building on the operational validation setup, we predict intensity change at a fixed lead time $\tau \in \{24, 48, 72, 96, 120\}$ h, defined as

$$\Delta V_\tau = V_{\max}(t + \tau) - V_{\max}(t). \tag{1}$$

Targeting ΔV_τ rather than $V_{\max}(t + \tau)$ is consistent with SHIPS and our experimental design: (i) the tropical system is known to exist at initialization and $V_{\max}(t)$ is observed, so $V_{\max}(t + \tau)$ follows once ΔV_τ is forecast; (ii) the target distribution is typically better behaved than the absolute intensity.

For each storm time t, we form a standardized predictor vector $\mathbf{x}_t \in \mathbb{R}^{p_{\tau,k}}$ composed of fold-specific causal features selected for the lead time τ (Section 3). We then learn a deterministic mapping

$$f_{\tau,k} : \mathbf{x}_t \mapsto \Delta V_\tau, \tag{2}$$

with one model (and potentially one feature set) per lead time τ and cross-validation fold k.

Parameters are estimated by least squares for consistency with SHIPS and with our linear, partial-correlation–based causal discovery. This yields a deterministic forecast without an explicit predictive distribution. We adopt this deterministic objective because (i) our focus is short lead times where mean-squared error is a standard operational target, (ii) the causal discovery framework deployed here is tailored to mean relationships rather than full conditional distributions, and (iii) maintaining compatibility with SHIPS facilitates a clean comparison in the baseline experiments that follow. Probabilistic forecasting at medium-range and subseasonal horizons is important but out of scope here.

3.5.2 Baseline Regression models and Baseline Feature Selection Methods

We compare the performance of PC-based causal feature selection models to different baselines to establish the validity of causal discovery methods to suggest useful predictors. Feature selection based on feature correlation and random forest feature importance are used as the chosen linear and nonlinear baselines to compare against the causal feature

selection outputs. The correlation baseline is obtained by ranking the correlation between the input variables and rate of TC intensity change. The variable ranking is then used to train linear regression models with increasing complexity in a sequential order. We follow the same process for the random-forest-based baseline, the main difference is that the variable ranking is now the Gini impurity-based feature importance of trained random forest regression models that predict =TC intensity change rates. We repeat these feature selection baselines for the withASSUMPS and noASSUMPS versions of the ERA5 and SHIPS-based experiments.

3.5.3 Regression architecture

All predictor sets are evaluated using simple multiple linear regression (MLR) models to maintain interpretability and isolate the effect of feature selection. For each experiment, we apply consistent 7-fold cross-validation settings and forecast lead times (24–120 hr), reporting performance on out-of-sample test sets to assess generalizability. We further evaluate the predictor sets using multilayer perceptron (MLP) models, which we select based on their ability to capture nonlinear behaviors in the data that cannot be expressed with MLR models. Each of the MLP models in our study includes a total of 5 layers, wherein each layer includes 512 units (also referred to as neurons). The first three layers rely on a Rectified Linear Unit (ReLU) activation function, the fourth layer instead uses hyperbolic tangent (tanh) to allow the model to output positive and negative intensity changes, and the final layer linearly combines the output of the 4th layer. All of our models were trained to minimize the mean square error (MSE) using the Adam optimizer [46] with default parameters, a learning rate of 0.001, and with early stopping set to end training if the validation loss is greater than the average of the last 50 epochs once the model has trained for at least 50 epochs. We note that this simplified architecture was chosen with rules of thumb and has not been subjected to a rigorous hyperparameter search given that the objective of these MLPs is not to achieve maximum performance, but rather to provide a simple nonlinear baseline to serve as a comparison to the MLRs.

4 Results

4.1 Effectiveness of multidata causal discovery for hurricane intensity prediction

We apply the multidata PC algorithm to ERA5 reanalysis data exclusively, avoiding inconsistencies from combining different observational sources. For the target variable—the intensity change over various lead times (24 to 120 hours)—we use observed maximum wind speed from IBTrACS to maintain consistency across experiments. We conduct two parallel sets of experiments for each lead time: (1) With SHIPS link assumptions (withASSUMPS): We encode prior knowledge by imposing direct causal links between operational SHIPS predictors and the intensity change target in the TiGraMITe M-PC algorithm. withASSUMPS respects the physical relationships of the original SHIPS model while allowing discovery of additional causal links among other variables. (2) Without SHIPS link assumptions (noASSUMPS or "kitchen-sink"): We allow the algorithm to freely explore causal links among all available ERA5 predictors without any predefined constraints. Figure 2 illustrates an example outcome for the 24-hour intensity change forecast (DELV24) from a representative cross-validation fold (Fold 3) under the kitchen-sink setup. The top panel shows training, validation, and test R^2 scores as a function of the number of selected predictors, controlled by the M-PC hyperparameter pc_α, which varies from $1.5 \ 10^{-4}$ to 0.6. Here, pc_α represents the significance level used in the conditional independence tests within the PC-stable algorithm. Higher values of pc_α correspond to a less strict statistical threshold for rejecting independence, allowing more predictors to be included in the causal model whereas a lower pc_α is more strict, resulting in a more stringent selection of predictors.

The bottom panel presents the variable selection abacus, highlighting predictor presence and groupings across the range of pc_α values. The encircled dots highlight the predictors included in the final shortlisted set, identified based on their consistent selection frequency across multiple folds. The same procedure was repeated for seven cross-validation folds, all forecast lead times, and both experimental setups. Given the volume of results, Figure 2 presents one representative fold without SHIPS link assumptions, while the best performing folds for lead times of up to 120 hours are provided in the Supplementary Information (Figures S1–S10). Predictors consistently selected across folds and validated configurations form a refined set that complements or improves upon the baseline SHIPS predictors for TC intensity change forecasting.

4.2 Causal feature selection outperforms baseline methods

Figure 3 summarizes the process and outcomes of identifying robust predictors using our causal feature selection pipeline for the 24-hour intensity change forecast (DELV24). Panel (a) shows how often each candidate predictor was selected in the seven cross-validation folds. By applying a clear threshold, which requires a predictor to appear in more than three folds—we filter out variables with inconsistent contributions and focus on predictors with stable links to
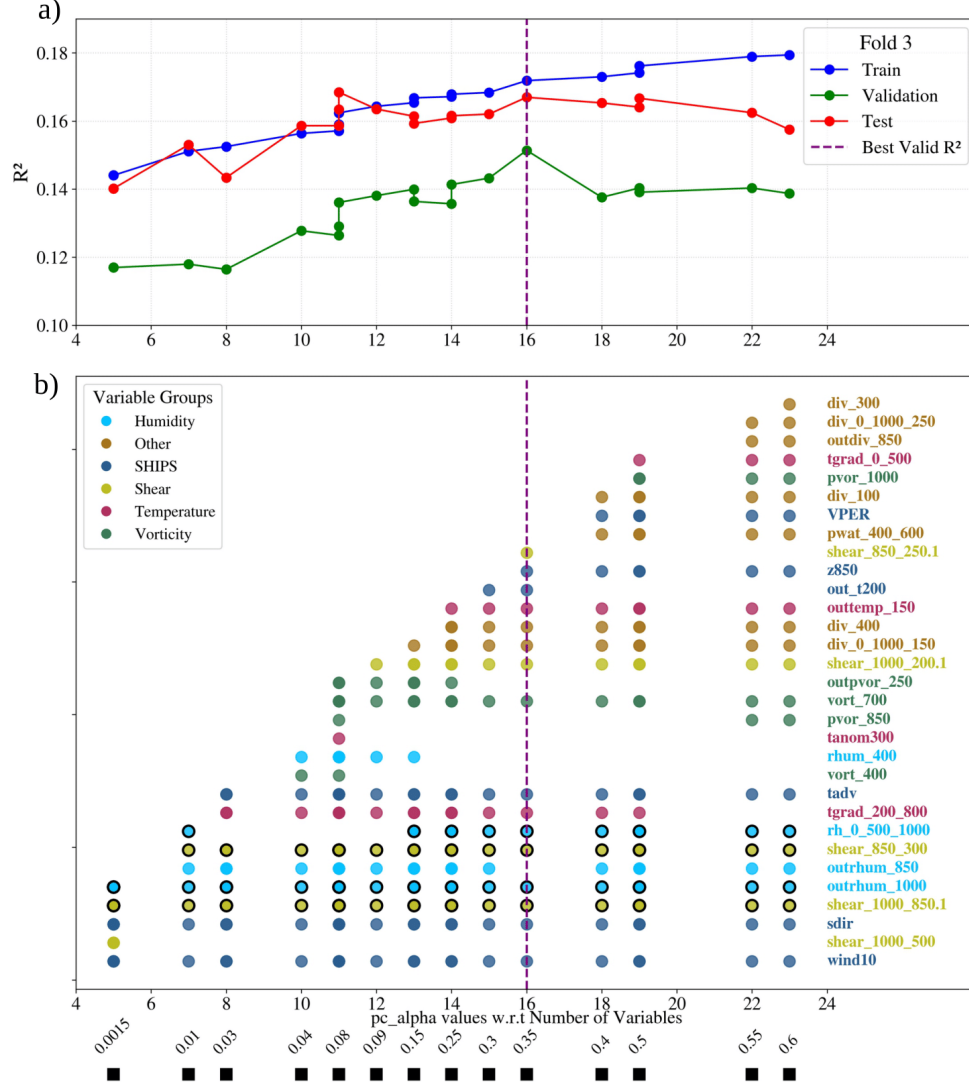
Figure 2: Example results for the 24-hour intensity change forecast (DELV24) from Fold 3 using the SHIPS+ERA5 predictor set for SHIPS predictors. (a) Coefficient of determination R^2 on training, validation, and test sets plotted against the number of selected predictors, each point corresponding to a different value of the M-PC causal discovery hyperparameter `pc_alpha` (bottom scale). The vertical dashed line indicates the configuration with the highest validation R^2. (b) Variable selection abacus: each dot shows the presence of a predictor across the `pc_alpha` range. Variables are colored by group (e.g., Original SHIPS predictors, Shear, Humidity), vertical dashed line marks the best validation score, and encircled dots highlight the occurrence of new shortlisted predictors for SHIPS.

intensity change. The color gradient highlights the strength of each variable's presence, from vivid cyan for consistently selected predictors to vivid pink for those with minimal presence. Panel (b) compares test R^2 scores for DELV24 across four feature selection strategies. Causal feature selection achieves the highest median skill, outperforming correlation ranking, random forest importance, and the no-selection baseline. This demonstrates that incorporating causally motivated predictors yields clear improvements in model generalizability. Comparing the causal and correlation feature selection methods, we observe that causal yields fewer outliers in prediction errors compared to the correlation-based approach, yet displays slightly larger uncertainty bounds. This behavior can be attributed to the fact that PC-based causal discovery method tends to identify stable, interventionally relevant relationships rather than purely predictive ones. When applied to a system with underlying nonlinear dynamics, the PC-based model may underfit complex regions of the predictor space, leading to broader prediction intervals, which reflects the model's caution in extrapolating beyond what the causal structure supports. This caution mitigates against large, spurious prediction errors, and results in fewer
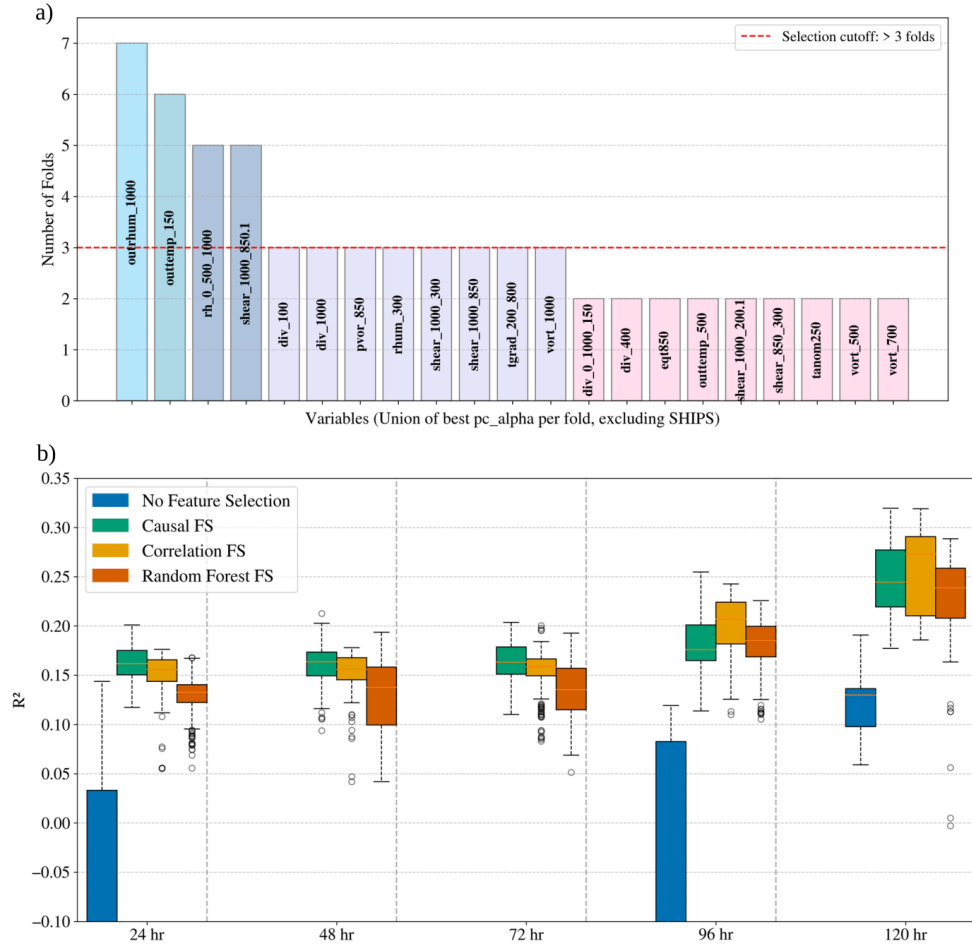
Figure 3: Summary of results for the 24-hour intensity change forecast (DELV24) using SHIPS+ERA5 predictors without SHIPS link assumptions. (a) Bar plot showing the frequency of each variable's selection across the best models from all seven cross-validation folds. A red dashed line marks the threshold (more than 3 folds) used to shortlist robust predictors for inclusion in the final SHIPS+ list. (b) Boxplot comparing test R² values for target DELV for each lead times 24, 48, 72, 96, 120 hrs for experiments with kitchen-sink approach (without link assumptions) across four feature selection strategies: causal discovery, correlation ranking, random forest importance, and no selection. Causal feature selection yields the highest median R² until 72 hrs lead time, showing improved generalization in a purely statistical prediction setup.

outliers and more robust predictions on unseen test cases. The trade-off between model stability and uncertainty shown here highlights the value of causal methods over correlation for building reliable statistical models for high-impact forecasting applications. By extending this comparison across all lead times (Fig S11, S12 in the SI) and evaluating the consistency of variable selection, we ultimately created a shortlist a new set of predictors to be added to SHIPS. These variables reliably contribute to improved forecast skill for short-range intensity prediction, particularly up to 72 hours. Beyond this time frame, the benefit of causal feature selection diminishes, emphasizing that while our method provides meaningful gains for in short-range, longer lead times will likely require integrating statistical models with dynamical forecast guidance to capture additional sources of predictability.

4.3 Recommending causally relevant predictors

In the previous section, we established that the causal predictors improve the generalizability of statistical TC intensity models for shorter lead times up to 72 hours but not for longer lead times (Fig. 3b).

Table 1: Recommended additional predictors to the operational SHIPS model.

Variable Group	Replication Code	Lead Time	SHIPS Code	Variable Description
Shear	Shear_1000_850	24h, 48h, 96h	SHL0	Vertical shear 1000–850 hPa, area-averaged 200–800 km.
	Shear_850_300	24h, 48h	SHMD	Vertical shear 850–300 hPa, area-averaged 200–800 km.
	Shear_1000_850.1	48h, 72h, 120h	SHL1	Vertical shear 1000–850 hPa, area-averaged 200–1000 km.
Humidity	Outrhum_1000	24h, 48h	R000	Relative Humidity at 1000 hPa, area-averaged 200–800 km.
	RH_0_500_1000	48h, 72h	R001	Relative Humidity at 1000 hPa, area-averaged 0–500 km.
Potential Vorticity	Outpvor_500	72h	PVOR	Potential Vorticity at 500 hPa, area-averaged 200–800 km.

Here, we use the coefficient of determination (R^2) as the primary performance metric, where positive values indicate that the model explains some fraction of the variance in TC intensity change, while negative values indicate that the model performs worse than simply predicting the mean. To increase the robustness of the predictor list and reduce the uncertainties arising from model hyperparameters and cross-validation strategies [47], the variable lists from the best MLR models for different folds are aggregated as a summary variable list. Of the different variables in the list, only those that are chosen in more than half of the cross-validation folds (more than 3 times) will be considered as candidate features to be included in the operational SHIPS for testing. Comparison of frequently repeated variables for experiments with and without SHIPS link assumptions (Fig. 3 a, Fig. S11, Fig. S12) yields a final shortlist of six final predictors that mostly describe lower and middle tropospheric vertical wind shear (e.g. SHL0, SHL1, SHMD), surface and boundary layer moisture (R000, R001), and midtropospheric potential vorticity conditions (PVOR) in the outer area of TC. The complete list of additional predictors recommended is provided in Table 1.

Before putting the discovered variables into operational models for testing, it is important to ensure that they pass the "expert judgment", as there is reasonable plausibility that these predictors could be causally related to TC intensity change. Low-tropospheric wind shear has been shown to be more negatively correlated with the intensity of Pacific typhoons that occurred in active typhoon seasons than commonly used deep-layer shear [48]. **(author?)** [49] further conclude that low-level shear produces quasiperiodic oscillations in the intensity of the TC that are related to the variability in the moisture of the boundary layer induced by the TC rainbands. Although TC intensification is generally most associated with relative humidity above the boundary layer [50], the moisture above the boundary layer reflects the balance between surface flux moistening and dry air ventilation near the surface [51]. The 1000 hPa relative humidity in the outer environment (R000), is a variable available in the SHIPS developmental file but unused in operational models. This shows that causal selection can reveal relevant predictors missed by traditional screening, improving short-range skill. This suggests that addition of overlooked factors like surface RH alongside mid-tropospheric humidity and shear helps capture multiscale processes drive intensity change, supporting the idea that boundary layer moisture could play a key role in regulating TC intensities [52] and the onset of rapid intensification [53]. Finally, mid-tropospheric vortex strength is critical in the early intensification of TCs [54, 55].

4.4 Added value of the nonlinear SHIPS+ models

The results presented in previous sections suggested six additional predictors causally related to the change in TC intensity that need to be tested in the same setup as the operational SHIPS, where the GFS analysis and forecast fields are used to derive predictors. As a first step towards a fair evaluation of operational model skill, we compare models trained with the original SHIPS predictors and the enriched SHIPS+, which includes the six newly identified causal predictors. For consistency with the ERA5 experiments, only the initial analysis time data are used to derive both SHIPS developmental and SHIPS+ datasets. This differs from the operational SHIPS setup, which incorporates dynamical model forecast output, so the results here may be negatively biased at longer lead times when synoptic-scale environmental changes are important.

Using R^2 as the performance metric, regression models trained with the additional causal predictors (yellow and orange boxes in Fig. 4) outperform models trained with only the original SHIPS predictors (green and blue boxes in Fig. 4) on unseen test TC cases, particularly at shorter lead times. This improvement is seen in both the linear MLR (blue and orange boxes) and nonlinear MLP (green and yellow boxes) models. However, the nonlinear SHIPS+ MLPs generally outperform their linear MLR counterparts. Since the SHIPS model is fully linear, we are interested in determining
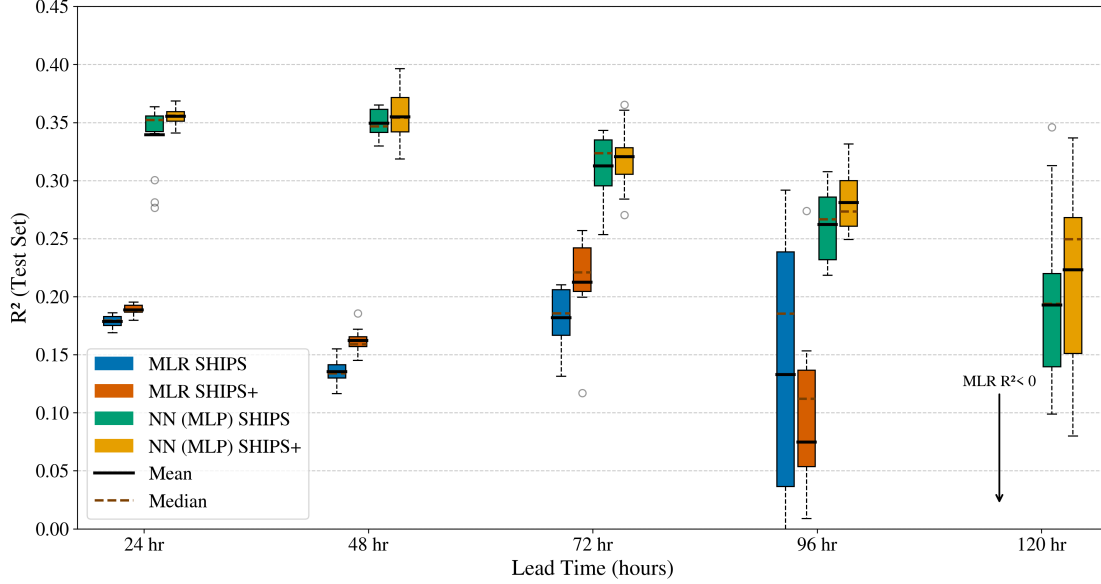
Figure 4: Comparison of test R^2 values across forecast lead times (24–120 hr) for the original SHIPS predictors (blue/green boxes) and the expanded SHIPS+ predictors (orange/yellow boxes), using no feature selection. Both MLR and MLP runs are shown to illustrate the added value of nonlinear modeling. Dashed brown lines indicate the median, and solid black lines mark the mean. Overall, the MLP consistently outperforms the MLR, demonstrating improved skill when nonlinearity is captured, while the inclusion of additional predictors in SHIPS+ further enhances forecast performance, especially at shorter lead times. Note that the SHIPS+ MLR R^2 drops below 0 at the 120 hr lead time, which is indicated in the figure by a downward arrow.

whether the superior MLP test skills can instead be attributed to the inability of MLRs to capture nonlinear relationships between the causal predictors and TC intensity changes. Indeed, the lead-time dependence of test skill improvements changes significantly in a nonlinear regression framework. The SHIPS+ MLRs show a clear lead-time dependence in the effect of the additional causal predictors, performing worse than the original SHIPS model beyond 72 hours, with negative R^2 at 120 hours. This lead time dependence is reduced in a nonlinear regression framework, where SHIPS+ outperforms SHIPS for all lead times. The results in Figure 4 implies that the relationship between the identified causal predictors and the rate of intensity changes of the TC can be approximated linearly at shorter lead times but not at extended lead times, which makes the MLRs incapable of utilizing the additional information beyond 72 hours. Finally, the wider uncertainty ranges of the trained models at longer lead times suggest that long-lead-time predictions cannot be adequately constrained with predictors derived solely from analysis time. This is a fundamental limitation of our predictive modeling setup, but it could be alleviated in operational settings by using dynamical model forecasts.

To better understand why the nonlinear SHIPS+ models outperform the other models, we conducted a SHAP (SHapley Additive exPlanations) analysis for the 24-hr lead time MLP and MLR models using both SHIPS and SHIPS+ datasets (see Fig. 5). SHAP values provide a model-agnostic interpretation of feature importance by quantifying the marginal contribution of each predictor to the model's output. To conduct the SHAP analysis, we rely on Kernel SHAP [56], using 300 samples from the training dataset to retrieve the background signal and evaluating on the 240 samples in the test set. Focusing on the differences between the linear MLR and the MLP models, we observe that of the six causal predictors, the MLP primarily utilizes the near-surface moisture predictors (red bars in Fig.5a), whereas the SHIPS+ MLRs (Fig.5b) use the causal shear predictors instead. A potential reason for MLPs and MLRs to select different causal predictors is that the relationship between near-surface moisture and short-term changes in TC intensity is too nonlinear to be used by linear MLRs. The dependencies of the causal boundary layer moisture predictors (e.g., R001) on the MLR predictions are nonlinear (Fig. 5e). Apart from prioritizing additional nonlinear dependencies, the SHIPS+ MLPs could also overperform their MLR counterparts by learning different dependencies for the existing SHIPS predictors. Figure 5d presents two examples in which this is the case, where low-level wind shear (SHL1) and mid-tropospheric potential vorticity (PVOR) have a much stronger dependence on MLP predictions than the muted ones for MLR models. Furthermore, MLPs learned a dependence between SHL1 and short-term intensity evolution that is strikingly different from the MLR dependence. Although intensification predictions in both the nonlinear and linear regression frameworks have a similar dependency on potential intensity (POT), which is the most critical predictor in both types of models,
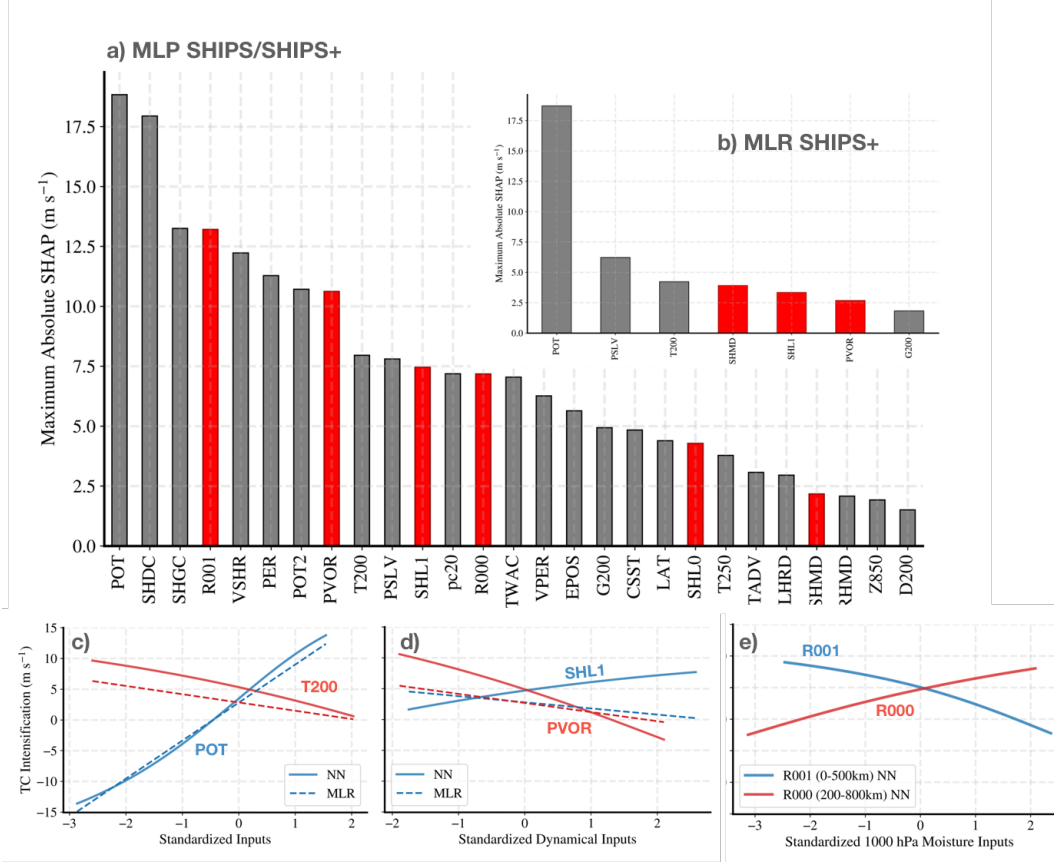
Figure 5: Predictor importance and dependencies for models trained on SHIPS+. (a–b) Global feature importance ranked by mean |SHAP| for the MLP (a) and MLR (b). (c) SHAP dependence for baseline SHIPS predictors POT (potential intensity minus current intensity) and T200 (200-hPa temperature, 200–800 km-averaged). (d) SHAP dependence for causally selected predictors SHL1 (1000–850-hPa vertical wind shear, 200–1000 km-averaged) and PVOR (500-hPa potential vorticity, 200–800 km-averaged). (e) For near-surface humidity, the MLP learns opposite dependencies: negative with R001 (1000-hPa relative humidity, 0–500 km-averaged) and positive with R000 (1000-hPa relative humidity, 200–800 km-averaged). All predictors in panels c-e are standardized.

the difference in the selection of other leading predictors and their respective learned dependencies contribute to the significant differences in the generalizability of the model to test TCs (Fig. 4).

4.5   Linear vs Nonlinear Models with Causal Predictors: Case Study

To illustrate the role of causally relevant predictors, we present Hurricane Larry (2021) as a case study.We selected Larry because, among the storms in the independent test set, it showed the largest increase in predictive skill (R^2) when using SHIPS+ MLPs with causal predictors compared to the operational SHIPS model. This allows us to investigate which features and causal predictors contributed most to the improvement in forecast performance. Larry was a long-lived and intense Cape Verde hurricane that underwent a period of rapid intensification over the tropical Atlantic before eventually making landfall in Newfoundland as a Category 1 hurricane. For this analysis, we truncate the time series before its extratropical transition, focusing solely on the tropical phase. Figure 6a) shows the best track of Larry during the evaluation period (Left), and a comparison of model performance (Right) between MLR and MLP, both with and without the addition of causally selected predictors. MLR SHIPS and MLR SHIPS+ both generally under-predict intensity change throughout the period, showing less variability and weaker response to fluctuations. Both versions of the Multi-Layer Perceptron (MLP) models with SHIPS and SHIPS+ (causally selected features) are consistently closer to the ground truth (black line) than the corresponding MLR models. MLP SHIPS initially over-predicts, but adding SHIPS+ dampens this overestimation, bringing it closer to the true signal, especially after 54 hours. In contrast, the performance gain from adding causal variables to MLR is minimal, indicating that linear models cannot fully leverage
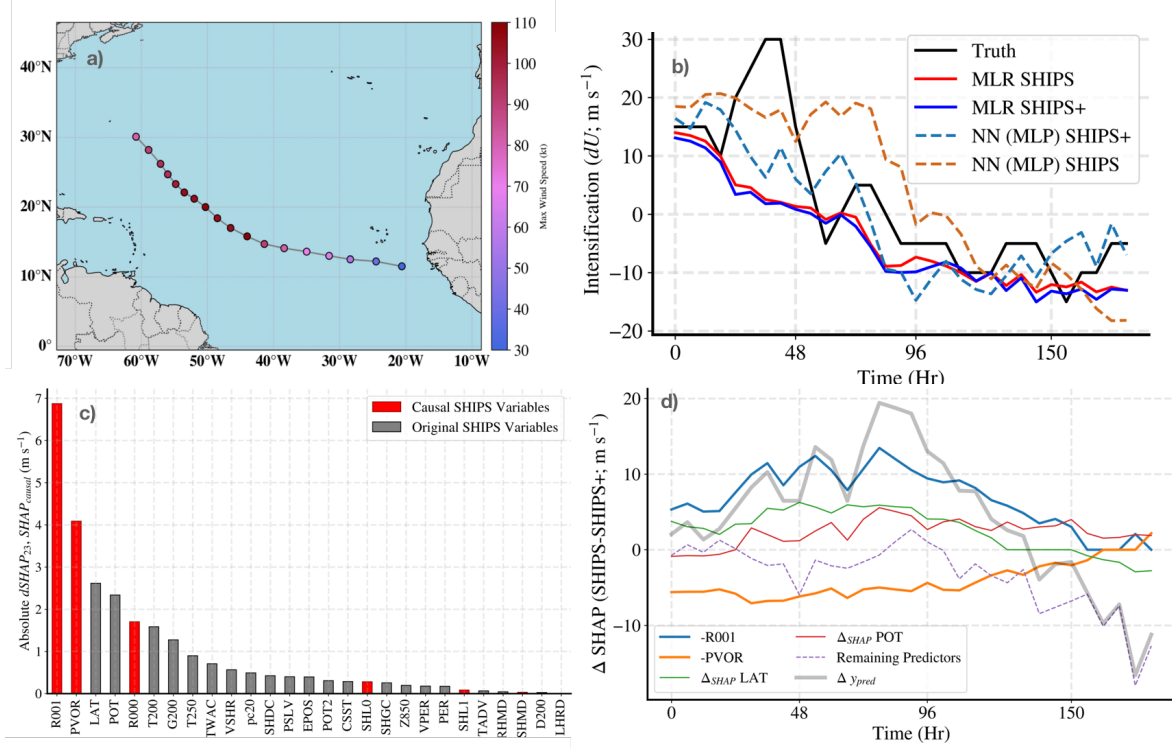
Figure 6: Hurricane Larry (2021): (a) Best track from IBTrACS showing the time used for testing (b) performance comparison of MLR vs. MLP comparing IBTrACS 24 hr intensity change (truth) to predictions with (SHIPS+) and without causal predictors (SHIPS). Overall, the MLP consistently outperforms the MLR, with the addition of predictors improving the performance notable from 54 hours. (c) Lifetime-maximum absolute SHAP values highlighting the dominant role of the added predictors (R001, PVOR). (d) Time-series decomposition of Δy_{pred} showing R001 and PVOR as the main contributors, with smaller adjustments from LAT and POT.

the additional causal features, highlighting the limitations of linear regression for dynamic systems like TCs, while the MLP models show a more dynamic response to intensity fluctuations.

To assess the role of causal predictors, we test whether the SHIPS+ MLP's gains arise from learning new relations for the added predictors or simply from reweighting existing SHIPS predictors. Therefore, we apply the SHAP-based decomposition of **(author?)** [57] (Appendix B) to compare the SHIPS model f and the SHIPS+ model g. By SHAP additivity, the prediction difference can be decomposed as follows:

$$\Delta y_{\text{pred}} = B_f - B_g + \underbrace{\sum_{i=1}^{N_{\text{SHIPS}}} \left(\text{SHAP}_i^f - \text{SHAP}_i^g \right)}_{\text{common predictors}}$$

$$- \underbrace{\sum_{i=1}^{N_{\text{causal}}} \text{SHAP}_i^g}_{\text{added (causal) predictors}}, \tag{3}$$

where B_f and B_g are the SHAP "background values" (mean training-set predictions) for SHIPS and SHIPS+, respectively. The first sum aggregates the change in SHAP contributions for the N_{SHIPS} predictors present in both models; the second subtracts the SHAP contributions of the N_{causal} predictors included only in SHIPS+.

For Hurricane Larry, lifetime-mean absolute SHAP values (Fig.6c) indicate that Eq.3 is well approximated by the contributions of two added predictors (R001 and PVOR) plus adjustments to two existing predictors (LAT and POT). The time-series decomposition (Fig.6d) confirms that R001 and PVOR dominate Δy_{pred}, with smaller adjustments from LAT and POT and a small residual from all other features over the period of interest (30-120 hr). This lends credence

to the interpretation that SHIPS+ improvements primarily arise from the added causal predictors. Consistent with the partial dependence for R001 (Fig.5e), SHIPS overestimates Larry's intensification because the original predictors under-represent the suppressing influence of inner-domain near-surface humidity; including R001 mitigates this bias.

4.6 Testing the new predictors in operational SHIPS

As described by [13], many new predictors have been added to SHIPS since the operational version was first implemented in 1991. As a preliminary test of how the research results presented above might improve the operational SHIPS model, the six potential new predictors listed in Table 1 were evaluated using the standard SHIPS procedures for annual updates. For this test, the North Atlantic data from 1982-2021 were used for training and the 2022-2024 cases were used for testing. This procedure does not use validation data because the SHIPS prediction coefficients are uniquely determined from the MLR fit to the training data. SHIPS uses GFS model fields so the predictors in Table 1 were recalculated from the GFS and added to the 28 predictors in the 2025 operational SHIPS model.

As in SHIPS development, candidate variables are added to the developmental set and undergo a predictor-screening test for statistical significance with respect to intensity changes at 6-h increments to lead times of 168 h. Training follows the SHIPS perfect-prognosis protocol: predictors are averaged over each forecast interval and sampled from GFS *analysis* fields about the future best-track positions at the valid times. In operations, the same predictors are computed from GFS *forecast* fields along the National Hurricane Center (NHC) forecast track. Note that the causal predictor identification used in this manuscript (Table 1) differs by using only $t = 0$ values about the current best-track position—by design, to keep feature discovery in a purely statistical forecasting (not post-processing) setting and to make the causal tests consistent with the regression task. We then apply the standard SHIPS screening and coefficient-update procedure described next. Three conditions are needed to pass the screening step: (1) When added one at a time, a new predictor must increase the variance explained by the model by at least 0.2% averaged over 5 consecutive forecast intervals; (2) The regression coefficient must be significant at the 99% level for at least 5 forecast intervals; (3) All predictors that pass steps 1 and 2 are added to the training sample and the coefficients are recalculated. The regression coefficients for each predictor must still be significant at the 99% level for at least five forecast intervals. If a predictor does not pass step 3, the least significant predictor is removed and then step 3 is repeated until all retained predictors pass the significance test. Results showed that SHMD, R001 and PVOR passed steps 1 and 2 for the 1982-2021 North Atlantic training sample. When all three were added, they also passed step 3. The next step in testing new SHIPS predictors is to perform retrospective forecasts on independent cases using only the input that is available in real time (forecast tracks and GFS forecast fields). To allow for a fair comparison, SHIPS was trained on the 1982-2021 sample with the original 28 operational predictors (baseline) and then with the addition of the three new predictors and both versions were run on the 2022-2024 independent cases. This sample included 794 cases with a 12 h forecast, which decreased to 114 cases by 168 h since many TCs dissipate by seven days. The mean absolute error (MAE) of the intensity forecasts for the baseline SHIPS and with the three new predictors were calculated using the intensity in the final NHC best track as "truth" following the standard NHC forecast procedure [58]. The verification sample includes tropical and subtropical cyclones but excludes the extratropical and pre-genesis stages.

Figure 7 shows the percent improvement (reduction in MAE) of SHIPS with the new predictors relative to baseline. SHIPS forecasts improved at all forecast times, with the largest improvements at longer forecast times. The improvements at 120-168 h were statistically significant at the 90% level using a standard statistical test that accounts for serial correlation. Although causal discovery used only $t = 0$ values, the longer-lead time gains seen in Figure 7 likely reflect the statistical–dynamical use of forecast fields in operations, motivating extension of the discovery procedure to predictors evaluated during the forecast period. The three new predictors (SHMD, R001 and PVOR) will be considered for implementation in future operational versions of SHIPS.

5 Conclusions

To improve operational SHIPS, we introduced a multidata PC causal discovery framework to discover causally relevant predictors that drive North Atlantic Hurricane intensity changes. First, we conducted tests by replicating the SHIPS predictors with the ERA5 reanalysis: we expanded the predictor list by testing combinations of dynamic and thermodynamic variables across different vertical layers and storm radial areas. M-PC tests using this augmented dataset revealed additional predictors that consistently demonstrated clear links to TC intensity change. Compared to other feature selection baselines, the causal approach performed best for short-range lead times (24–72 hours), highlighting its ability to isolate physically meaningful predictors while filtering out spurious relationships. Through this process, we shortlisted six key variables that were not currently used as operational SHIPS predictors. One of the new predictors, the surface level relative humidity for the outer area of the storm (R000), stands out because it is (1) already part of the extended SHIPS predictors list; and (2) well known to be critical for predicting rapid intensification.
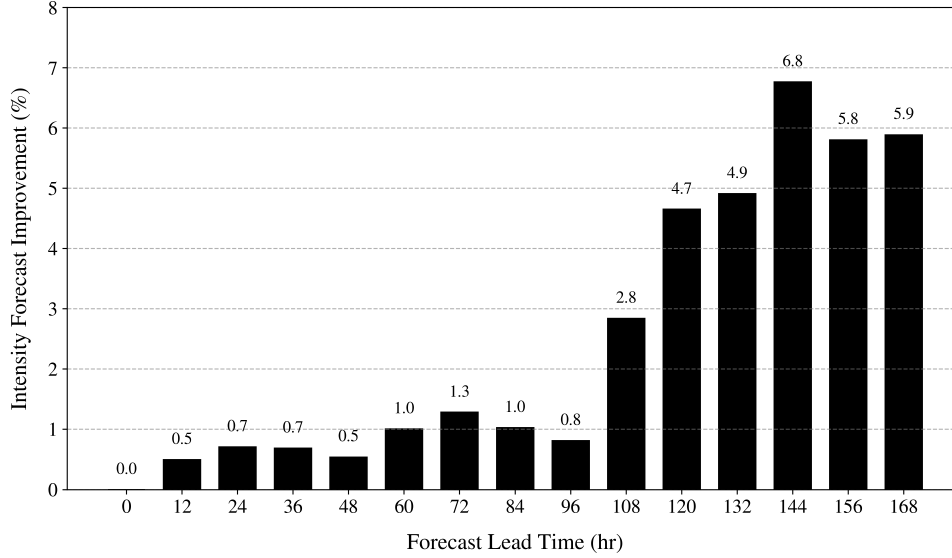
Figure 7: The improvement in the SHIPS intensity forecasts with the three new predictors relative to baseline for independent 2022–2024 Atlantic basin forecasts with real-time input.

The inclusion of other predictors, such as low-level shear and mid-tropospheric potential vorticity, underscores the importance of multiscale dynamical interactions and moisture–vorticity coupling processes that are often missed by traditional statistical frameworks.

This expanded predictors set, termed "SHIPS+", was rigorously tested by cross-validation across multiple folds and lead times. When using GFS analysis-time data from the SHIPS developmental dataset, incorporating these causal predictors consistently improved the forecast skill for short-range lead times up to 72 hours. The comparison of MLR and nonlinear MLP models further revealed that while linear models capture much of the intensity changes at shorter lead times, nonlinear interactions become increasingly important as lead time increases. The MLP consistently outperforms the MLR for all lead times and shows improved skill for SHIPS+ over the original SHIPS baseline, underscoring the value of combining causal feature selection with non-linear regression to better represent the evolving dynamics of TC intensification. This result suggests that the operational SHIPS model could be improved by replacing the MLR with a nonlinear method such as MLP in future versions.

The Hurricane Larry (2021) case study illustrated the added value of causally relevant predictors. Although both MLR and MLP models benefit from causal features, the nonlinear MLP better leverages these predictors to capture dynamic intensity fluctuations, including rapid intensification and decay phases, even though the predictors were selected using a *linear* causal discovery procedure. In contrast, linear models tend to under-predict variability and respond less effectively to short-term changes, highlighting the advantage of nonlinear approaches in capturing complex TC behavior. These findings are consistent with the operational SHIPS tests, showing improvements at all lead times and the largest MAE reductions at 120–168h (Fig.7).

While purely statistical models already benefit from causal feature selection and nonlinear methods, our results also point to their limitations at longer lead times, where static predictors become less informative. However, in a statistical-dynamical framework such as operational SHIPS, where the GFS forecast model provides the evolving environmental fields, these limitations can diminish, as evidenced by the largest improvements at 120–168 h. A promising next step is to derive causally selected predictors directly from dynamical forecast output and along forecast trajectories. This would allow their seamless integration into statistical-dynamical frameworks such as SHIPS, potentially improving the forecast skill for rapid intensification and addressing one of the main sources of uncertainty in operational TC intensity prediction.

Acknowledgments

Data availability

The causal discovery package and tutorials are freely available in the Tigramite GitHub repository. All the scripts, tutorials to replicate the experiments and preprocessed data for this study are available at Causal-SHIPS and have been archived in Zenodo. Large datasets not included in the GitHub repository for the Part 1 and Part 3 scripts are available at https://doi.org/10.5281/zenodo.17241222. The North Atlantic TC data are from the IBtrACS data archive. ERA5 datasets, including multiple and single pressure levels, were obtained from the Copernicus Climate Data Store at cds.climate.copernicus.eu. TC-PRIMED Data and documentation are available at (this link). The SHIPS developmental dataset can be accessed at NESDIS SHIPS Developmental Data. All datasets are publicly available and can be freely accessed and reused under their respective data policies.

References

[1] William M Gray. The formation of tropical cyclones. *Meteorology and atmospheric physics*, 67(1):37–69, 1998.

[2] RL Elsberry. Recent advancements in dynamical tropical cyclone track predictions. *Meteorology and Atmospheric Physics*, 56(1):81–99, 1995.

[3] Russell L Elsberry. Advances in research and forecasting of tropical cyclones from 1963–2013. *Asia-Pacific Journal of Atmospheric Sciences*, 50:3–16, 2014.

[4] Roger A Pielke Jr and Christopher W Landsea. Normalized hurricane damages in the united states: 1925–95. *Weather and forecasting*, 13(3):621–631, 1998.

[5] Kristen M. Crossett, Thomas J. Culliton, Peter C. Wiley, and Timothy R. Goodspeed. Population trends along the coastal united states, 1980–2008. Technical report, United States National Ocean Service, Special Projects, 2004.

[6] Suzana J Camargo, Joanne Camp, Russell L Elsberry, Paul A Gregory, Philip J Klotzbach, Carl J Schreck III, Adam H Sobel, Michael J Ventrice, Frédéric Vitart, Zhuo Wang, et al. Tropical cyclone prediction on subseasonal time-scales. *Tropical Cyclone Research and Review*, 8(3):150–165, 2019.

[7] Julian T Heming, Fernando Prates, Morris A Bender, Rebecca Bowyer, John Cangialosi, Phillippe Caroff, Thomas Coleman, James D Doyle, Anumeha Dube, Ghislain Faure, et al. Review of recent progress in tropical cyclone track forecasting and expression of uncertainties. *Tropical Cyclone Research and Review*, 8(4):181–218, 2019.

[8] Brian H Tang, Juan Fang, Alicia Bentley, Gerard Kilroy, Masuo Nakano, Myung-Sook Park, VPM Rajasree, Zhuo Wang, Allison A Wing, and Liguang Wu. Recent advances in research on tropical cyclogenesis. *Tropical Cyclone Research and Review*, 9(2):87–105, 2020.

[9] Mark DeMaria and John Kaplan. An updated statistical hurricane intensity prediction scheme (ships) for the atlantic and eastern north pacific basins. *Weather and Forecasting*, 14(3):326–337, 1999.

[10] Yu-qing Wang and C-C Wu. Current understanding of tropical cyclone structure and intensity changes–a review. *Meteorology and Atmospheric Physics*, 87(4):257–278, 2004.

[11] Mark DeMaria, James L Franklin, Matthew J Onderlinde, and John Kaplan. Operational forecasting of tropical cyclone rapid intensification at the national hurricane center. *Atmosphere*, 12(6):683, 2021.

[12] Kerry Emanuel and Fuqing Zhang. On the predictability and error sources of tropical cyclone intensity forecasts. *Journal of the Atmospheric Sciences*, 73(9):3739–3747, 2016.

[13] Mark DeMaria, James L Franklin, Rachel Zelinsky, David A Zelinsky, Matthew J Onderlinde, John A Knaff, Stephanie N Stevenson, John Kaplan, Kate D Musgrave, Galina Chirokova, et al. The national hurricane center tropical cyclone model guidance suite. *Weather and Forecasting*, 37(11):2141–2159, 2022.

[14] Yuqing Wang. Recent research progress on tropical cyclone structure and intensity. *Tropical Cyclone Research and Review*, 1(2):254–275, 2012.

[15] Ryan D Torn and Mark DeMaria. Validation of ensemble-based probabilistic tropical cyclone intensity change. *Atmosphere*, 12(3):373, 2021.

[16] Mark DeMaria and John Kaplan. A statistical hurricane intensity prediction scheme (ships) for the atlantic basin. *Weather and Forecasting*, 9(2):209–220, 1994.

[17] F. D. Marks and M. DeMaria. Development of a tropical cyclone rainfall climatology and persistence (r-cliper) model. Technical report, NOAA/OAR/AOML/Hurricane Research Division, 2003.

[18] John P Cangialosi, Eric Blake, Mark DeMaria, Andrew Penny, Andrew Latto, Edward Rappaport, and Vijay Tallapragada. Recent progress in tropical cyclone intensity forecasting at the national hurricane center. *Weather and Forecasting*, 35(5):1913–1922, 2020.

[19] Mark DeMaria, Michelle Mainelli, Lynn K Shay, John A Knaff, and John Kaplan. Further improvements to the statistical hurricane intensity prediction scheme (ships). *Weather and Forecasting*, 20(4):531–543, 2005.

[20] Jakob Runge. Causal network reconstruction from time series: From theoretical assumptions to practical estimation. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 28(7):075310, 2018.

[21] Jakob Runge, Sebastian Bathiany, Erik Bollt, Gustau Camps-Valls, Dim Coumou, Ethan Deyle, Clark Glymour, Marlene Kretschmer, Miguel D Mahecha, Jordi Muñoz-Marí, et al. Inferring causation from time series in earth system sciences. *Nature communications*, 10(1):1–13, 2019.

[22] Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science advances*, 5(11):eaau4996, 2019.

[23] Andreas Gerhardus and Jakob Runge. Causal discovery in ensembles of climate time series. Technical report, Copernicus Meetings, 2022.

[24] Jakob Runge, Andreas Gerhardus, Gherardo Varando, Veronika Eyring, and Gustau Camps-Valls. Causal inference for time series. *Nature Reviews Earth & Environment*, 4(7):487–505, 2023.

[25] Imme Ebert-Uphoff and Yi Deng. Causal discovery for climate research using graphical models. *Journal of Climate*, 25(17):5648–5665, 2012.

[26] Jakob Runge, Vladimir Petoukhov, and Jürgen Kurths. Quantifying the strength and delay of climatic interactions: The ambiguities of cross correlation and a novel measure based on graphical models. *Journal of climate*, 27(2):720–739, 2014.

[27] Lucile Ricard, Tom Beucler, Claudia Christine Stephan, and Athanasios Nenes. A causal intercomparison framework unravels precipitation drivers in global storm-resolving models. *npj Climate and Atmospheric Science*, 8(1):245, 2025.

[28] Marie C McGraw and Elizabeth A Barnes. Memory matters: A case for granger causality in climate variability studies. *Journal of climate*, 31(8):3289–3300, 2018.

[29] Jakob Runge, Vladimir Petoukhov, and Jürgen Kurths. Quantifying the strength and delay of climatic interactions: The ambiguities of cross correlation and a novel measure based on graphical models. *Journal of Climate*, 27(2):720 – 739, 2014.

[30] Elizabeth A Barnes and James A Screen. The impact of arctic warming on the midlatitude jet-stream: Can it? has it? will it? *Wiley Interdisciplinary Reviews: Climate Change*, 6(3):277–286, 2015.

[31] Jasper S Wijnands, Guoqi Qian, and Yuriy Kuleshov. Variable selection for tropical cyclogenesis predictive modeling. *Monthly Weather Review*, 144(12):4605–4619, 2016.

[32] Jakob Runge. Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. In Jonas Peters and David Sontag, editors, *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 of *Proceedings of Machine Learning Research*, pages 1388–1397. PMLR, 03–06 Aug 2020.

[33] B. Latos, I.-J. Moon, and D. H. Kim. Advancing seasonal hurricane predictions using causal ai. In *EGU General Assembly 2024*, Vienna, Austria, Apr 2024.

[34] H Hersbach, B Bell, P Berrisford, G Biavati, A Horányi, J Muñoz Sabater, J Nicolas, C Peubey, R Radu, I Rozum, et al. Era5 hourly data on pressure levels from 1979 to present, copernicus climate change service (c3s) climate data store (cds), 2018.

[35] Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020.

[36] Kenneth R Knapp, Michael C Kruk, David H Levinson, Howard J Diamond, and Charles J Neumann. The international best track archive for climate stewardship (ibtracs) unifying tropical cyclone data. *Bulletin of the American Meteorological Society*, 91(3):363–376, 2010.

[37] John A Knaff, Daniel P Brown, Joe Courtney, Gregory M Gallina, and John L Beven. An evaluation of dvorak technique–based tropical cyclone intensity estimates. *Weather and Forecasting*, 25(5):1362–1379, 2010.

[38] Muhammad Naufal Razin et al. Tropical cyclone precipitation, infrared, microwave, and environmental dataset (tc primed). *Bulletin of the American Meteorological Society*, 104(11):E1980–E1998, 2023.

[39] S. S Ganesh, Tom Beucler, Frederick Iat-Hin Tam, Milton S Gomez, Jakob Runge, and Andreas Gerhardus. Selecting robust features for machine learning applications using multidata causal discovery. *arXiv e-prints*, pages arXiv–2304, 2023.

[40] Marlene Kretschmer, Dim Coumou, Jonathan F Donges, and Jakob Runge. Using causal effect networks to analyze different arctic drivers of midlatitude winter circulation. *Journal of climate*, 29(11):4069–4081, 2016.

[41] Marlene Kretschmer, Judah Cohen, Vivien Matthias, Jakob Runge, and Dim Coumou. The different stratospheric influence on cold-extremes in Eurasia and North America. *npj Climate and Atmospheric Science*, 1(1), nov 22 2018.

[42] S.M. Samarasinghe, C. Connolly, E.A. Barnes, I. Ebert-Uphoff, and L. Sun. Strengthened causal connections between the MJO and the North Atlantic with climate warming. *Geophys. Res. Lett.*, 48:e2020GL091168, 2021.

[43] Fernando Iglesias-Suarez, Pierre Gentine, Breixo Solino-Fernandez, Tom Beucler, Michael Pritchard, Jakob Runge, and Veronika Eyring. Causally-informed deep learning to improve climate models and projections. *Journal of Geophysical Research: Atmospheres*, 129(4):e2023JD039202, 2024.

[44] Jakob Runge, Reik V Donner, and Jürgen Kurths. Optimal model-free prediction from multivariate time series. *Physical Review E*, 91(5):052909, 2015.

[45] Andrew R Lawrence, Marcus Kaiser, Rui Sampaio, and Maksim Sipos. Data generating process to evaluate causal discovery techniques for time series data. *arXiv preprint arXiv:2104.08043*, 2021.

[46] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[47] Lily-belle Sweet, Christoph Müller, Mohit Anand, and Jakob Zscheischler. Cross-validation strategy impacts the performance and interpretation of machine learning models. *Artificial Intelligence for the Earth Systems*, 2(4):e230026, 2023.

[48] Yuqing Wang, Yunjie Rao, Zhe-Min Tan, and Daria Schönemann. A statistical analysis of the effects of vertical wind shear on tropical cyclone intensity change over the western north pacific. *Monthly Weather Review*, 143(9):3434–3453, 2015.

[49] Hao Fu, Yuqing Wang, Michael Riemer, and Qingqing Li. Effect of unidirectional vertical wind shear on tropical cyclone intensity change—lower-layer shear versus upper-layer shear. *Journal of Geophysical Research: Atmospheres*, 124(12):6265–6282, 2019.

[50] Longtao Wu, Hui Su, Robert G Fovell, Bin Wang, Janice T Shen, Brian H Kahn, Svetla M Hristova-Veleva, Bjorn H Lambrigtsen, Eric J Fetzer, and Jonathan H Jiang. Relationship of environmental relative humidity with north atlantic tropical cyclone intensity and intensification rate. *Geophysical research letters*, 39(20), 2012.

[51] Michael S Fischer, Paul D Reasor, Brian H Tang, Kristen L Corbosiero, Ryan D Torn, and Xiaomin Chen. A tale of two vortex evolutions: Using a high-resolution ensemble to assess the impacts of ventilation on a tropical cyclone rapid intensification event. *Monthly weather review*, 151(1):297–320, 2023.

[52] Joshua B Wadler, David S Nolan, Jun A Zhang, and Lynn K Shay. Thermodynamic characteristics of downdrafts in tropical cyclones as seen in idealized simulations of different intensities. *Journal of the Atmospheric Sciences*, 78(11):3503–3524, 2021.

[53] Xiaomin Chen, Jian-Feng Gu, Jun A Zhang, Frank D Marks, Robert F Rogers, and Joseph J Cione. Boundary layer recovery and precipitation symmetrization preceding rapid intensification of tropical cyclones under shear. *Journal of the Atmospheric Sciences*, 78(5):1523–1544, 2021.

[54] Marja Bister and Kerry A Emanuel. The genesis of hurricane guillermo: Texmex analyses and a modeling study. *Monthly weather review*, 125(10):2662–2682, 1997.

[55] Bolei Yang and Zhe-Min Tan. Interactive radiation accelerates the intensification of the midlevel vortex for tropical cyclogenesis. *Journal of the Atmospheric Sciences*, 77(12):4051–4065, 2020.

[56] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[57] Arthur Grundner, Tom Beucler, Pierre Gentine, Fernando Iglesias-Suarez, Marco A. Giorgetta, and Veronika Eyring. Deep learning based cloud cover parameterization for icon. *Journal of Advances in Modeling Earth Systems*, 14(12):e2021MS002959, 2022. e2021MS002959 2021MS002959.

[58] John P. Cangialosi, Brad J. Reinhart, and Jonathan Martinez. National hurricane center forecast verification report: 2023 hurricane season. Technical Report Verification Report 2023, National Hurricane Center, NOAA, Miami, FL, June 2024. Published June 3, 2024; accessed August 30, 2025.

Supplementary Material: Multidata Causal Discovery for Statistical Hurricane Intensity Forecasting

Saranya Ganesh S., Frederick Iat-Hin Tam, Milton S. Gomez, Tom Beucler
Faculty of Geosciences and Environment,
Expertise Center for Climate Extremes,
University of Lausanne
Lausanne, Vaud, Switzerland

Marie McGraw, Mark DeMaria, Kate Musgrave
Cooperative Institute for Research in the Atmosphere,
Colorado State University,
Fort Collins, Colorado

Jakob Runge
Department of Computer Science,
University of Potsdam
Potsdam, Germany

This Supplementary Information (SI) provides additional details and results supporting the main manuscript.

- The full set of variables used as predictors in our tropical cyclone intensity prediction experiments, including the core SHIPS predictors, is provided in Tables S1–S5: Table S1 lists the original SHIPS predictors, Table S2 their replication from ERA5, Table S3 the additional inner-core variables (0–2°), Table S4 the outer-core variables (200–800 km, up to 1000 km), and Table S5 the extended set of predictors from the TC PRIMED dataset. The SHIPS replication experiment uses a subset of predictors consistent with those available from ERA5 reanalysis and the TC PRIMED dataset, excluding variables such as PC20, PSLV, SST, and SHGC, except when directly compared with operational SHIPS runs.

- Figures S1–S10, which are similar in format to Figure 2 in the main text, but display the best-performing validation fold (based on R^2) out of the seven experiments, with and without link assumptions, across the five lead times (24, 48, 72, 96, and 120 hours).

- Figures S11–S12, which are similar in format to Figure 3a in the main text, showing the frequency of each variable across the best models from all seven cross-validation folds for the experiments with and without SHIPS link assumptions for the target DELV at lead times from 24 to 120 hours. A cutoff of 3 is applied, where predictors are shortlisted if they appear at least 4 times across the 7 cross-validation experiments.

- Figures S13–S14, which are similar in format to Figure 3b in the main text, showing box plots of coefficient of determination (R^2) values comparing different feature selection methods for Train (Top), Validation (Middle), and Test (Bottom) for DELV at lead times from 24 to 120 hours, with and without SHIPS link assumptions.

- Figure S15, which is similar in format to Figure 4 in the main text, showing box plots of coefficient of determination (R^2) values comparing different feature selection methods for the SHIPS+ dataset for Training (Top) and Validation (Bottom) sets at lead times from 24 to 120 hours, with and without SHIPS link assumptions.

Table S1: SHIPS Developmental Dataset Variables

Variable	Description
DELV24/48/72/96/120	Target: Max surface wind difference over 24/48/72/96/120 h
PMIN	Minimum central pressure of the system (hPa)
VMAX	Maximum wind speed (kt)
PER	Intensity change (kt) over prior 12 h
VPER	PER multiplied by VMAX
PC20	Percent of GOES IR pixels colder than –20°C, area-averaged 50–200 km
SPDX	X-component of storm translational speed (kt) at forecast time
PSLV	Steering layer pressure center of mass (hPa) at forecast time
SST	Sea surface temperature at storm center (°C), time-averaged 0–48 h
POT	Potential intensity minus current intensity (kt), time-averaged 0–48 h
SHDC	850–200 hPa vertical wind shear magnitude (kt), time-averaged 0–48 h
T200	200 hPa temperature (°C), 200–800 km area-avg, time-avg 0–48 h
T250	250 hPa temperature (°C), background-subtracted
EPOS	Parcel instability parameter from equivalent potential temperature (°C)
RHMD	Relative humidity (%) in 500–700 hPa, 200–800 km area-avg
TWAT	Time tendency of average tangential wind within 500 km
Z850	Relative vorticity (10^{-7} s^{-1}), 0–1000 km area-avg
D200	200 hPa divergence (10^{-7} s^{-1}), 0–1000 km area-avg
LHRD	SHDC multiplied by sine of latitude
VSHR	VMAX multiplied by SHDC
POT2	Square of POT
SHGC	Generalized shear (kt) from 100–1000 hPa levels
SDIR	Deviation of 850–200 hPa shear direction from optimal (°)
TADV	Temperature advection between 850–700 hPa, area-avg 0–500 km
G200	Temperature perturbation (°C) at 200 hPa, 200–800 km, time-avg 0–48 h
LAT	Latitude of TC center

Table S2: ERA5 Replication of SHIPS Predictors

Variable	Description
DELV	Intensity change over 24/48/72/96 h
PMIN	Minimum central pressure (hPa)
VMAX	Max wind speed from ERA5 10m winds (kt)
PER	Intensity change (kt) over prior 12 h
VPER	PER multiplied by VMAX
SPDX	ERA5 x-component of storm translational speed (kt)
PSLV	ERA5 steering layer pressure center (hPa)
SST	ERA5 sea surface temperature (°C)
POT	ERA5 potential intensity minus current (kt)
SHDC	ERA5 850–200 hPa vertical shear (kt)
T200	ERA5 200 hPa temperature (°C)
T250	ERA5 250 hPa temperature (°C)
EPOS	ERA5 parcel instability from equivalent potential temperature (°C)
RHMD	ERA5 relative humidity (%) in 500–700 hPa
TWAT	ERA5 time tendency of average tangential wind
Z850	ERA5 850 hPa relative vorticity (10^{-7} s^{-1})
D200	ERA5 200 hPa divergence (10^{-7} s^{-1})
LHRD	ERA5 shear magnitude (kt × 10) with vortex removed
VSHR	ERA5 VMAX × SHDC (kt^2)
POT2	ERA5 POT squared (kt^2)
SHGC	ERA5 generalized shear (kt)
SDIR	ERA5 shear direction deviation (°)
TADV	ERA5 temperature advection (°C)

Table S3: Inner Core (0–200 km area-averaged) Variables from ERA5

Variable	Description	Pressure Levels (hPa)
div	Horizontal divergence (s^{-1})	100, 200, 250, 300, 400, 500, 700, 850, 1000
eqt	Equivalent potential temperature (K)	1000, 200, 250, 300, 400, 500, 700, 850
vort	Relative vorticity (s^{-1})	100, 150, 200, 250, 300, 400, 500, 700, 850, 1000
pvor	Potential vorticity (PVU)	100, 150, 200, 250, 300, 400, 500, 700, 850, 1000
rhum	Relative humidity (%)	100, 150, 200, 250, 300, 400, 500, 700, 850, 1000
gpot	Geopotential height (m)	100, 150, 200, 250, 300, 400, 500, 700, 850, 1000
temp	Air temperature (K)	100, 150, 200, 250, 300, 400, 500, 700, 850, 1000

Table S4: Outer Area (200–800 km area-averaged) Variables from ERA5

Variable	Description	Pressure Levels (hPa)
outdiv	Horizontal divergence (s^{-1})	100, 200, 250, 300, 400, 500, 700, 850, 1000
outeqt	Equivalent potential temperature (K)	1000, 200, 250, 300, 400, 500, 700, 850
outvort	Relative vorticity (s^{-1})	100, 150, 200, 250, 300, 400, 500, 700, 850, 1000
outpvor	Potential vorticity (PVU)	100, 150, 200, 250, 300, 400, 500, 700, 850, 1000
outrhum	Relative humidity (%)	100, 150, 200, 250, 300, 400, 500, 700, 850, 1000
outgpot	Geopotential height (m)	100, 150, 200, 250, 300, 400, 500, 700, 850, 1000
outtemp	Air temperature (K)	100, 150, 200, 250, 300, 400, 500, 700, 850, 1000

Table S5: TC PRIMED Variables used in this study.

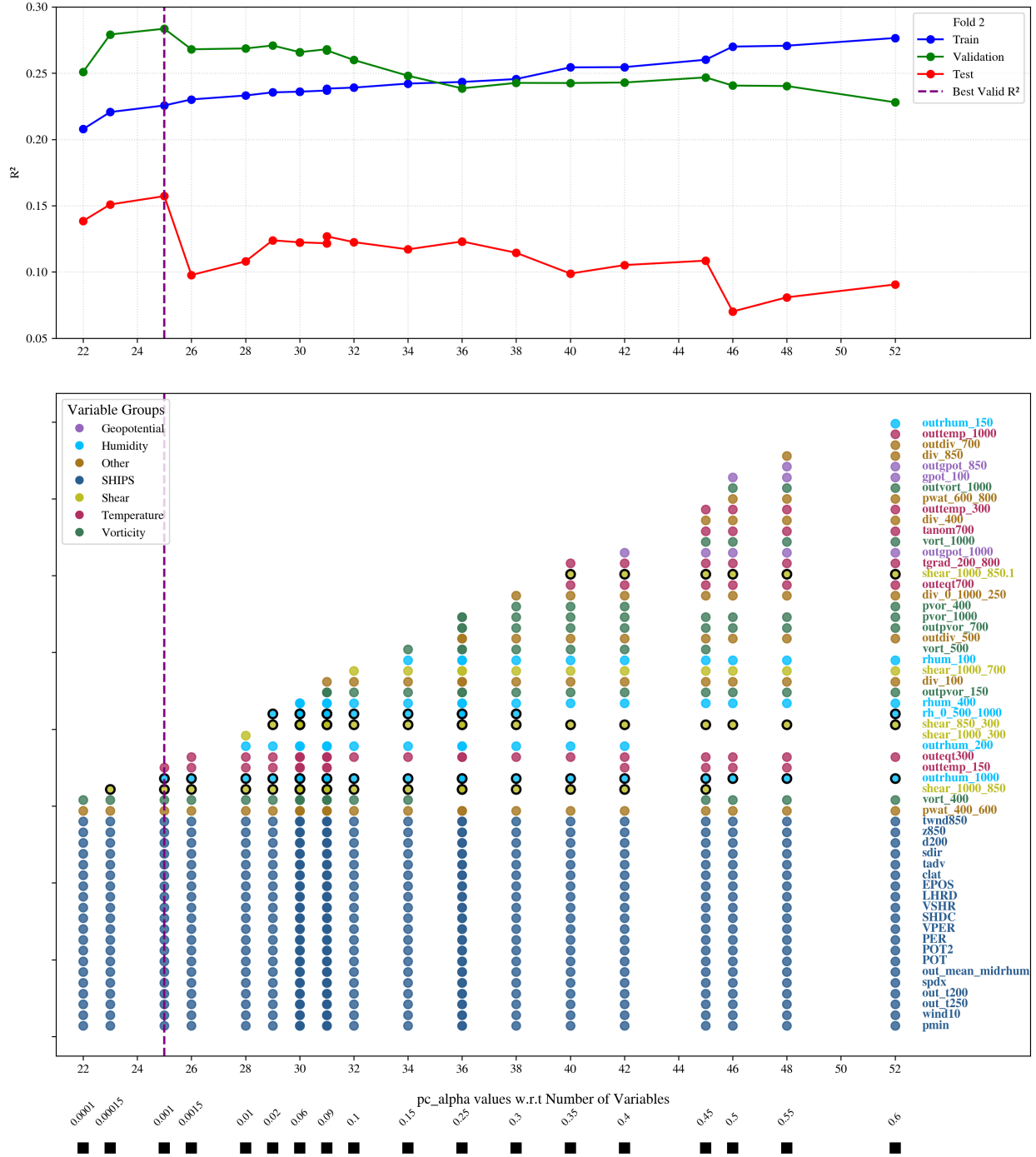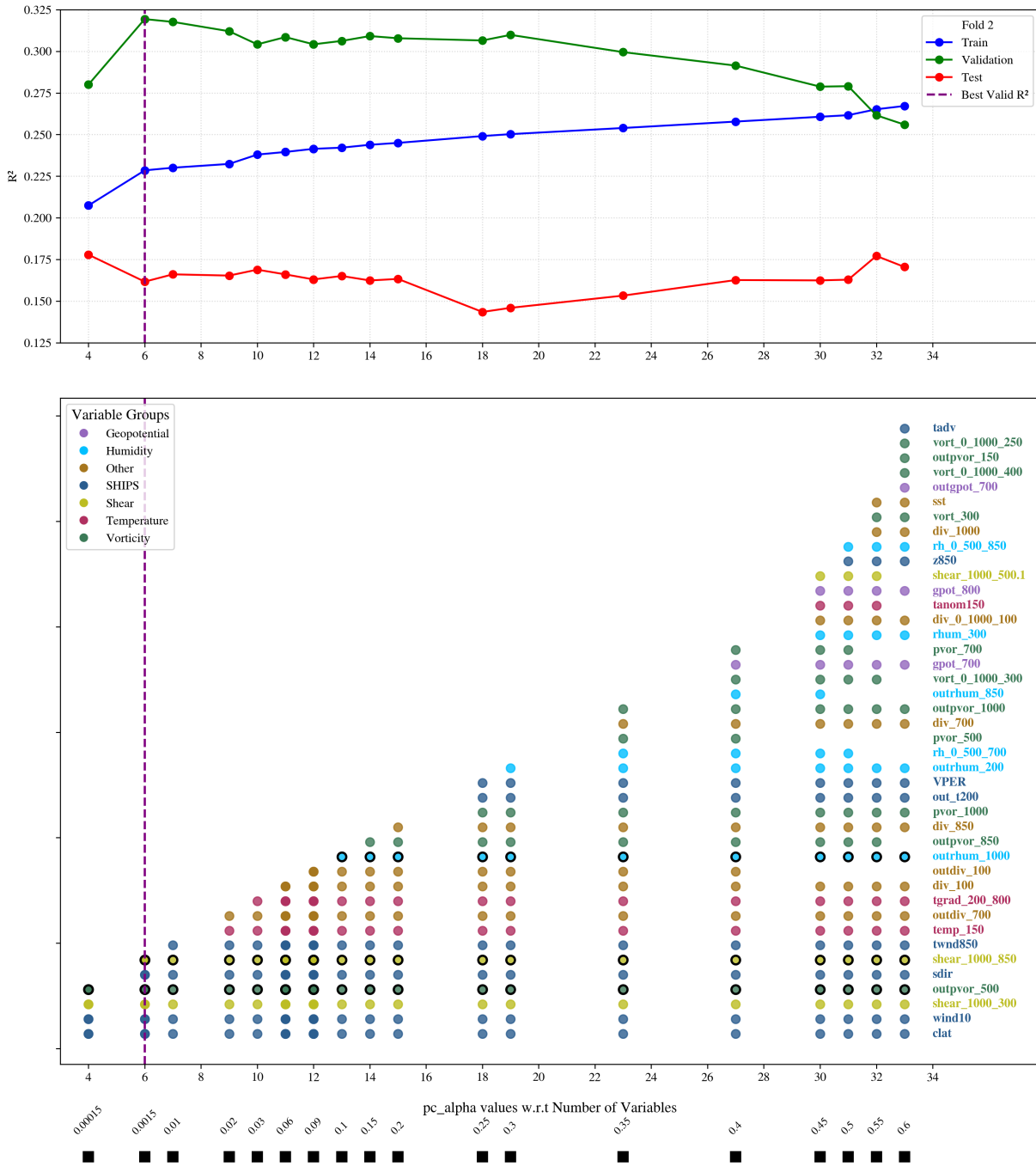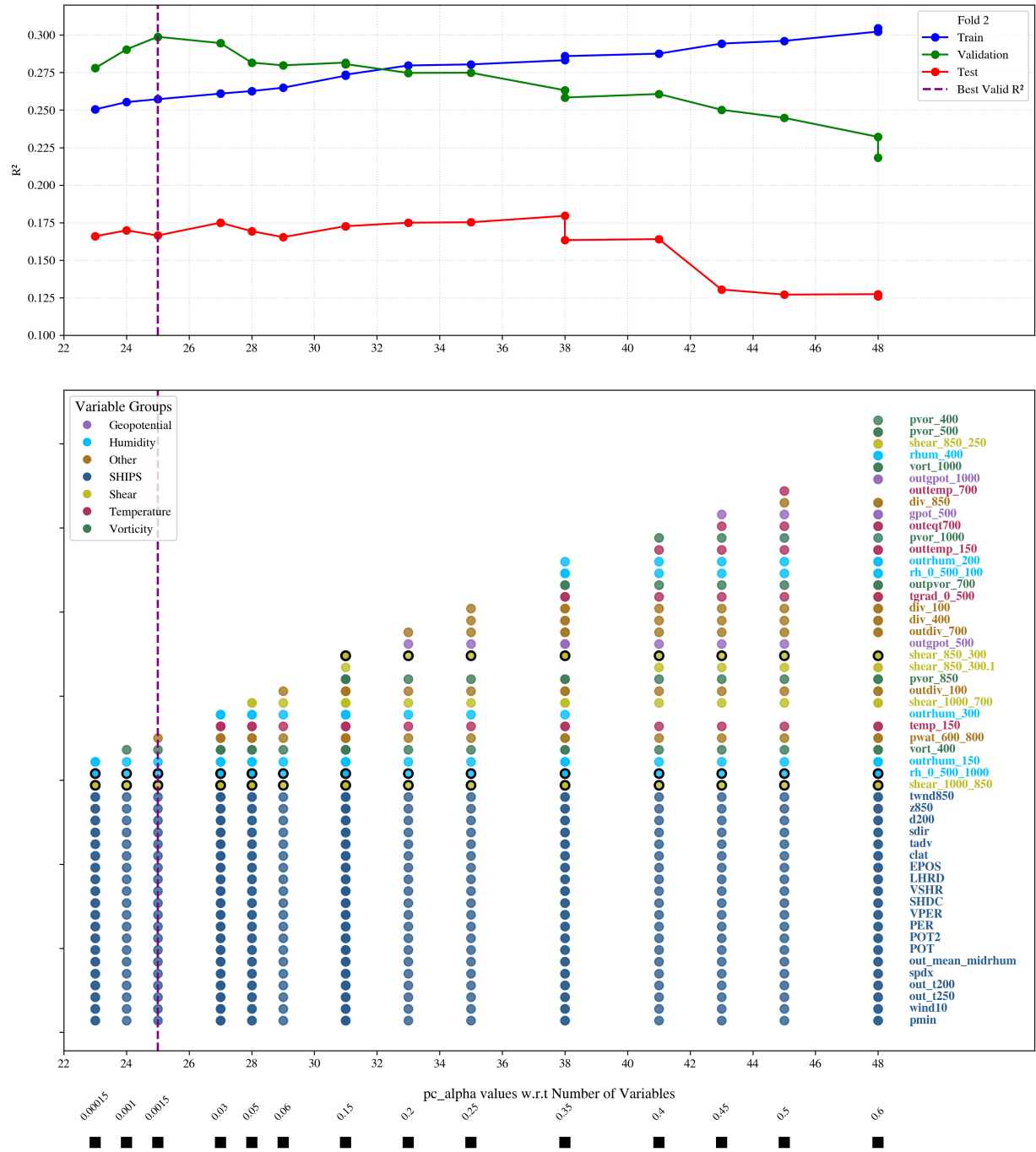| Variable	Description	Units	Radius (km)	Pressure Levels (hPa)
shear / shear.1	Vertical wind shear	m s^{-1}	200–800, 200–1000	850–200, 1000–300, 1000–500, 1000–700, 1000–850, 850–250, 850–300, 850–500
tgrad	Temperature gradient	K km^{-1}	0–500, 200–800	–
pwat	Precipitable water	mm	0–200, 200–400, 400–600, 600–800, 800–1000	–
div	Divergence	s^{-1}	0–1000	100, 150, 200, 250, 300, 400, 500, 700, 850, 1000
vort	Vorticity	s^{-1}	0–1000	100, 150, 200, 250, 300, 400, 500, 700, 850, 1000
geop	Geopotential height	m	0–1000	100, 150, 200, 250, 300, 400, 500, 700, 850, 1000
rh	Relative humidity	%	0–500	100, 150, 200, 250, 300, 400, 500, 700, 850, 1000
tanom	Warm-core temperature anomaly	K	0–15 km to 1500 km	100, 150, 200, 250, 300, 400, 500, 700, 850, 1000

Figure S1: Results for the 24-hour intensity change forecast (DELV24) from the best fold using the SHIPS+ERA5 predictor set *without* SHIPS link assumptions. **Top panel:** R^2 scores on training, validation, and test sets plotted against the number of selected variables, each point corresponding to a different value of the M-PC1 causal discovery hyperparameter `pc_alpha` (bottom scale). The vertical dashed line indicates the configuration with the highest validation R^2. **Bottom panel:** Variables selection: each dot shows the presence of a predictor across the `pc_alpha` range. Variables are colored by group (e.g., SHIPS, Shear, Humidity).

Figure S2: Results for the 24-hour intensity change forecast (DELV24) from the best fold using the SHIPS+ERA5 predictor set *with* SHIPS link assumptions. Same as the Figure S1.
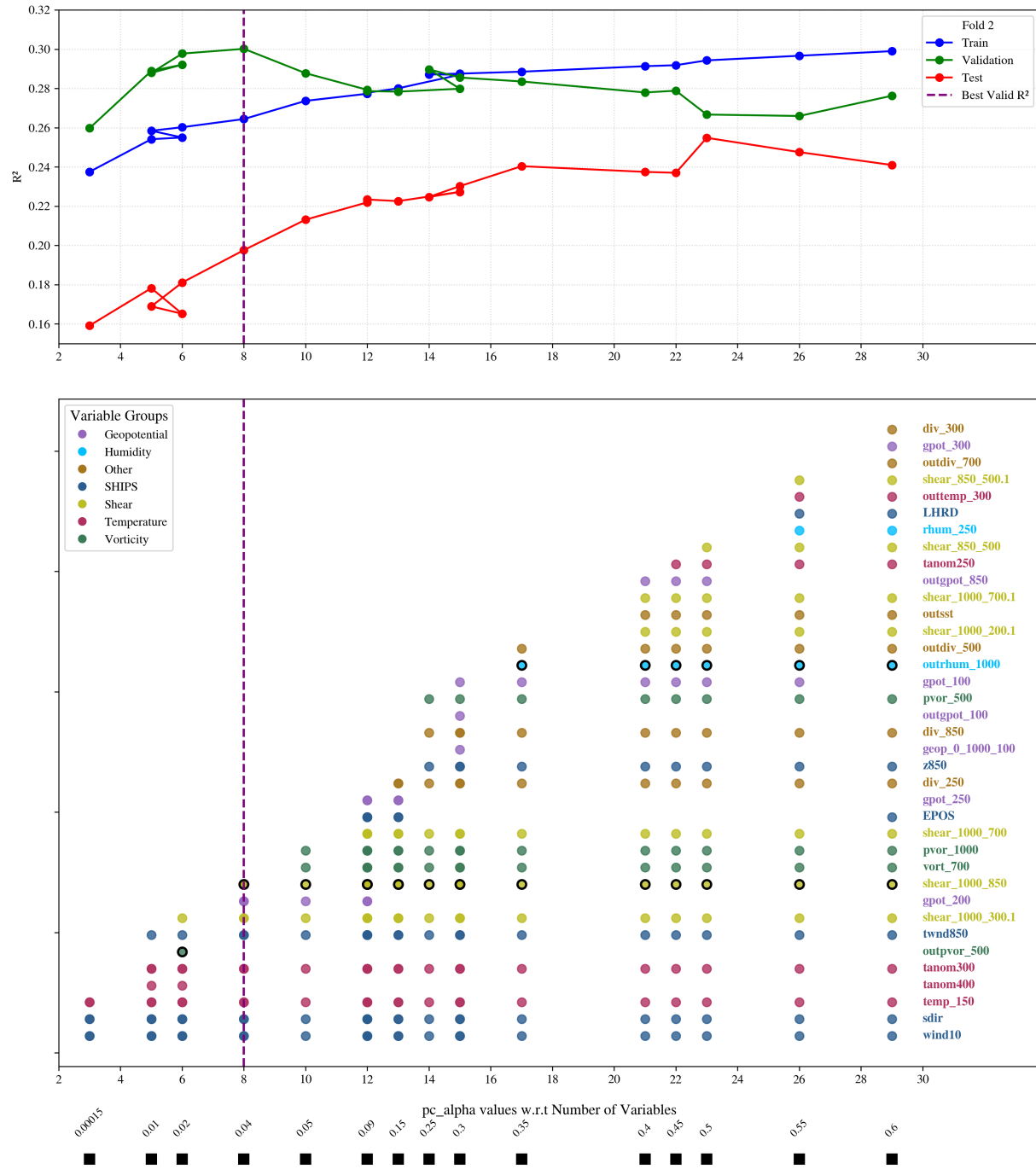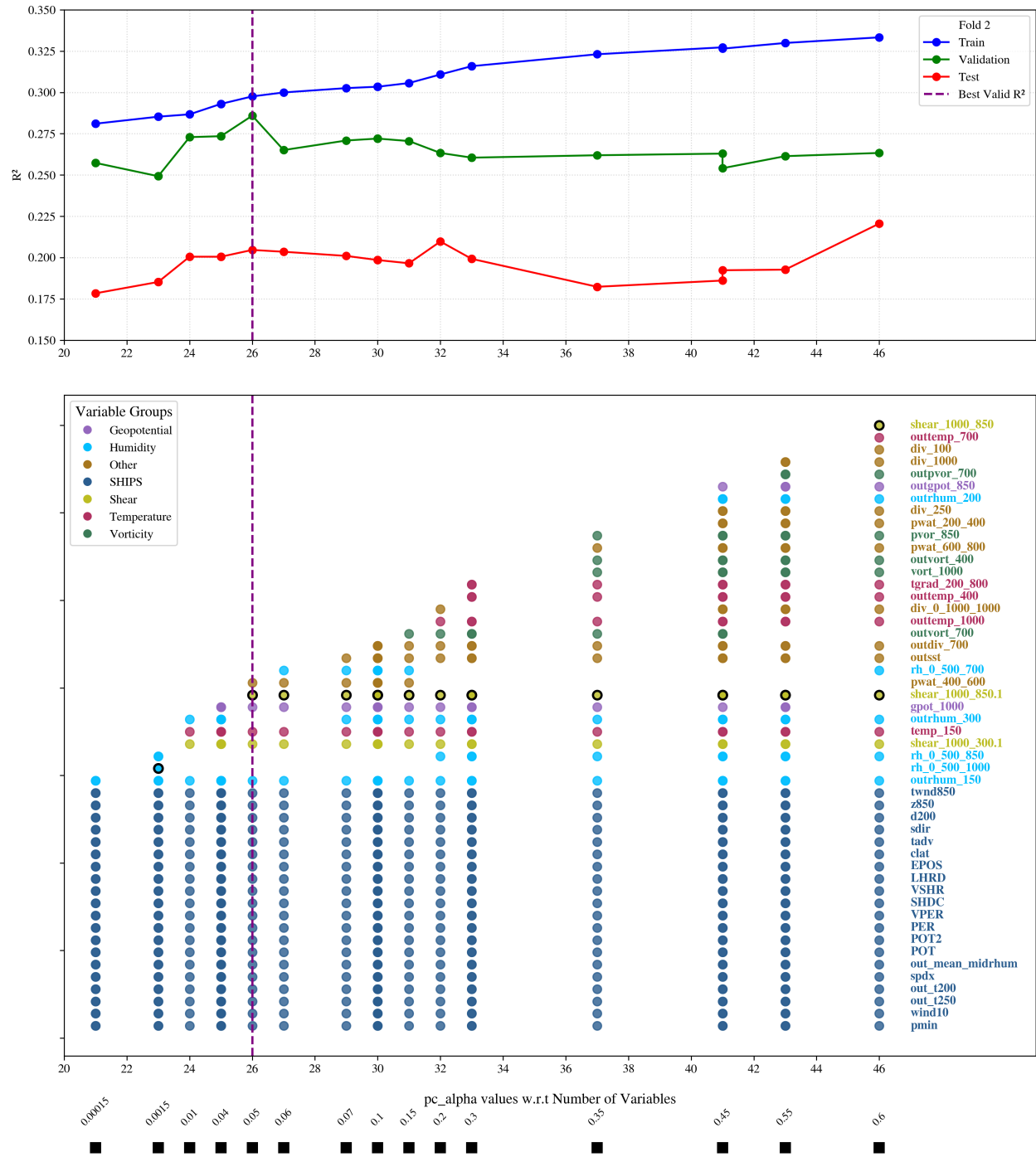
Figure S3: Results for the 48-hour intensity change forecast (DELV48) from the best fold using the SHIPS+ERA5 predictor set *without* SHIPS link assumptions. Same as the Figure S1.

Figure S4: Results for the 48-hour intensity change forecast (DELV48) from the best fold using the SHIPS+ERA5 predictor set *with* SHIPS link assumptions. Same as the Figure S2.
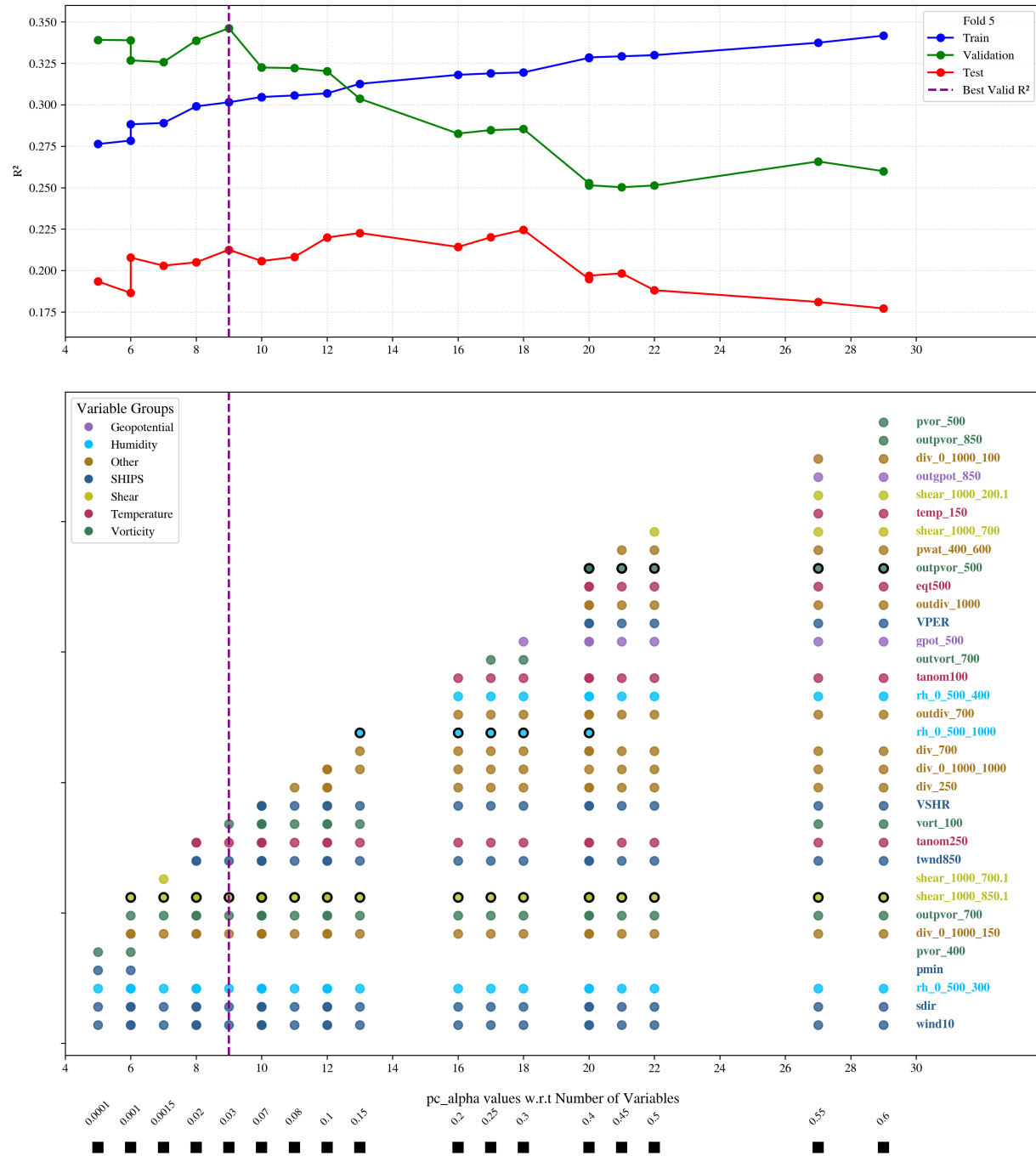
Figure S5: Results for the 72-hour intensity change forecast (DELV72) from the best fold using the SHIPS+ERA5 predictor set *without* SHIPS link assumptions. Same as the Figure S1.
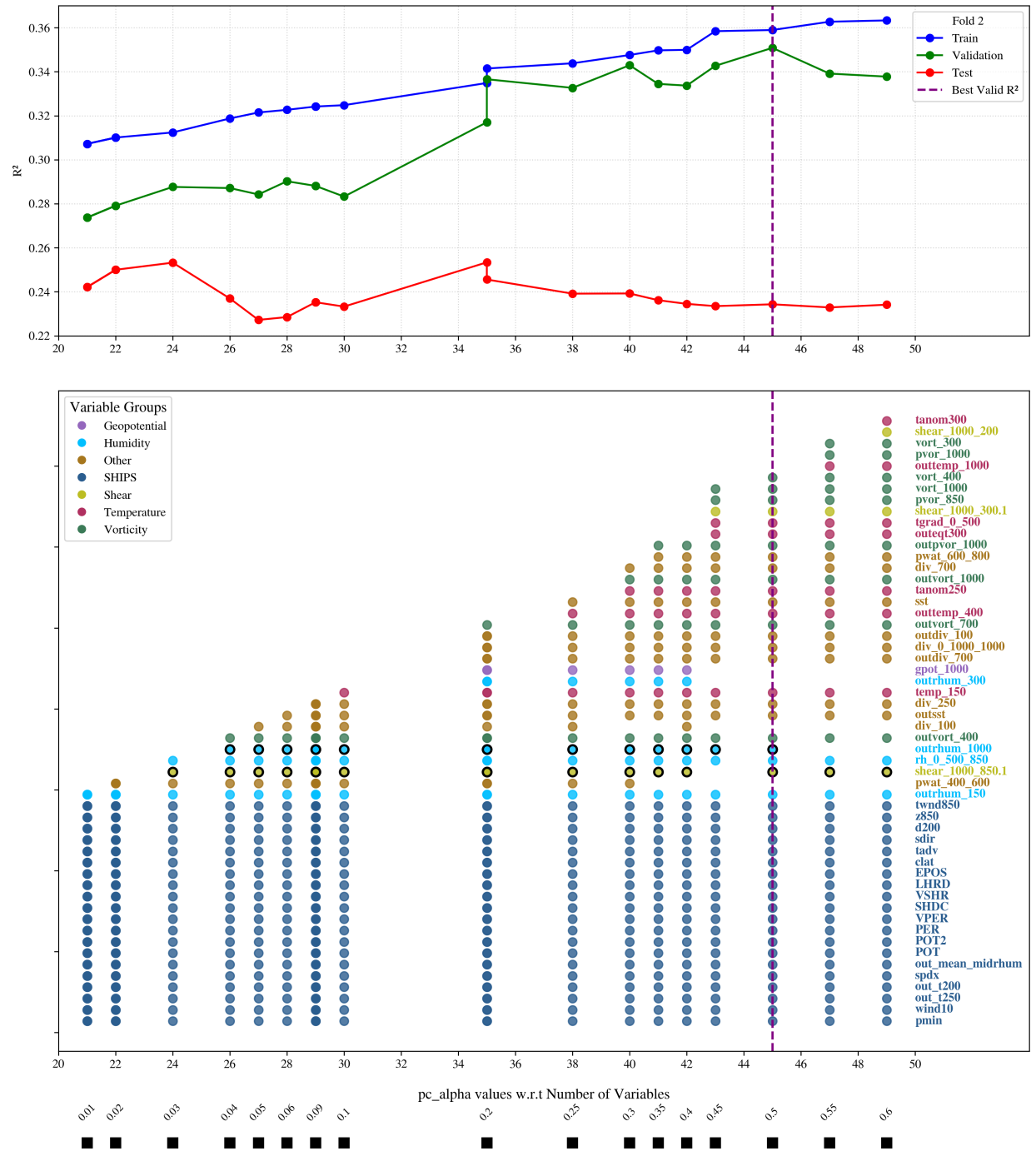
Figure S6: Results for the 72-hour intensity change forecast (DELV72) from the best fold using the SHIPS+ERA5 predictor set *with* SHIPS link assumptions. Same as the Figure S2.
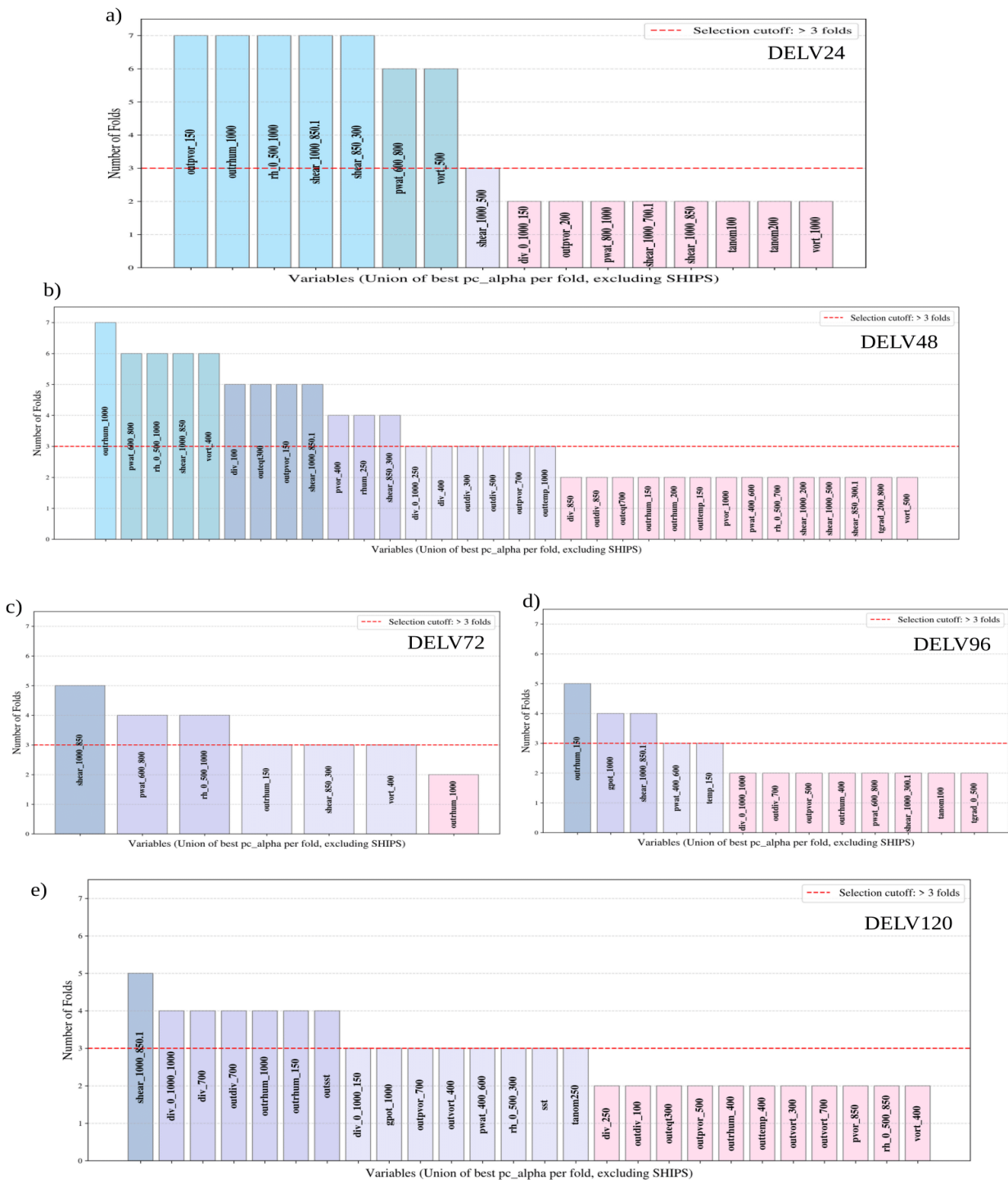
Figure S7: Results for the 96-hour intensity change forecast (DELV96) from the best fold using the SHIPS+ERA5 predictor set *without* SHIPS link assumptions. Same as the Figure S1.

Figure S8: Results for the 96-hour intensity change forecast (DELV96) from the best fold using the SHIPS+ERA5 predictor set *with* SHIPS link assumptions. Same as the Figure S2.

Figure S9: Results for the 120-hour intensity change forecast (DELV120) from the best fold using the SHIPS+ERA5 predictor set *without* SHIPS link assumptions. Same as the Figure S1.

Figure S10: Results for the 120-hour intensity change forecast (DELV120) from the best fold using the SHIPS+ERA5 predictor set *with* SHIPS link assumptions. Same as the Figure S2.

Figure S11: Bar plots showing the frequency of each variable across the best models from all seven cross-validation folds for experiments with SHIPS Link assumptions for target DELV for lead times 24 hrs to 120 hrs. Red dotted line shows the cut off for variable shortlist.
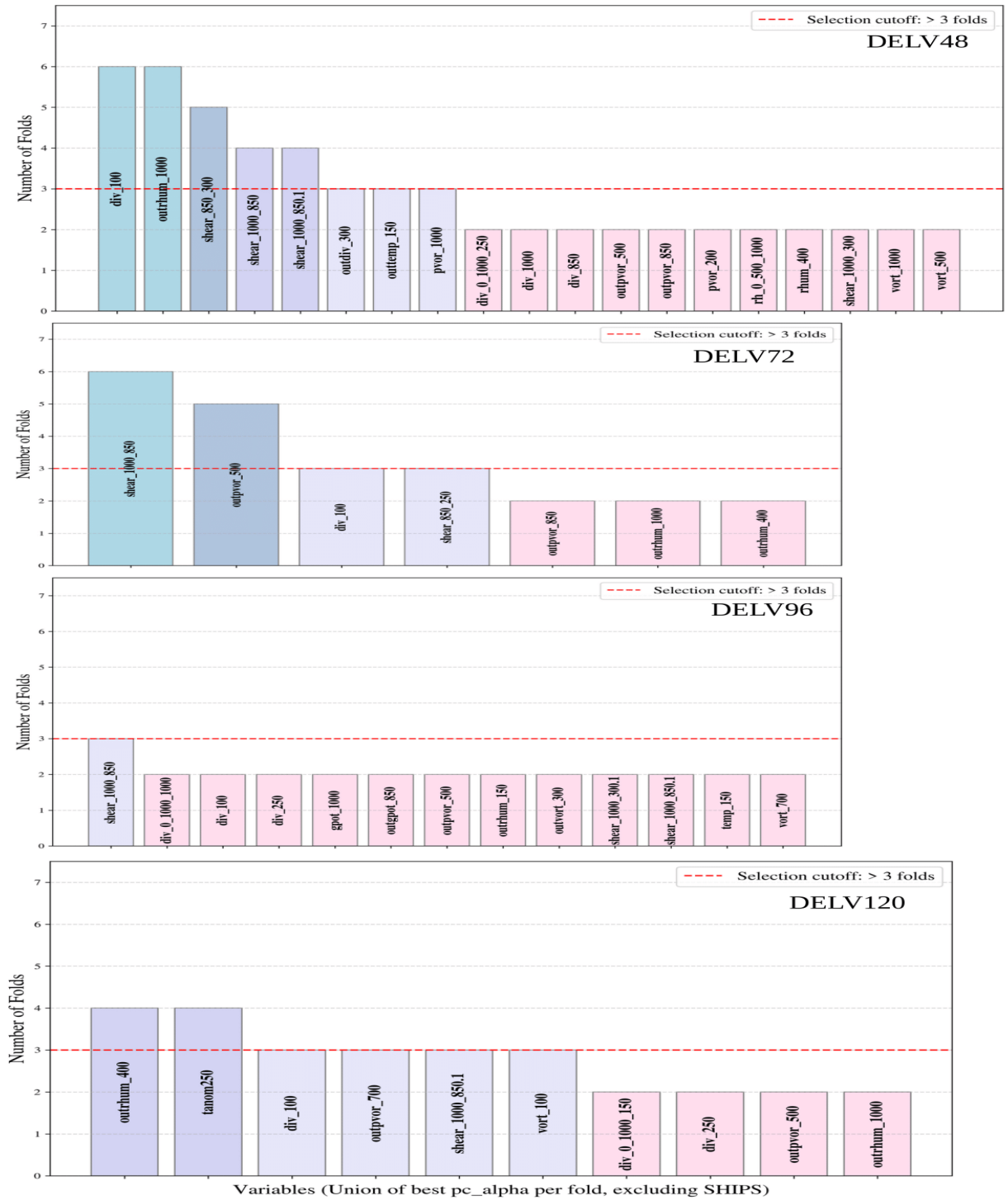
Figure S12: Bar plots showing the frequency of each variable across the best models from all seven cross-validation folds for experiments without any SHIPS link assumptions for target DELV for lead time 48 hrs to 120 hrs. Red dotted line shows the cut off for variable shortlist.
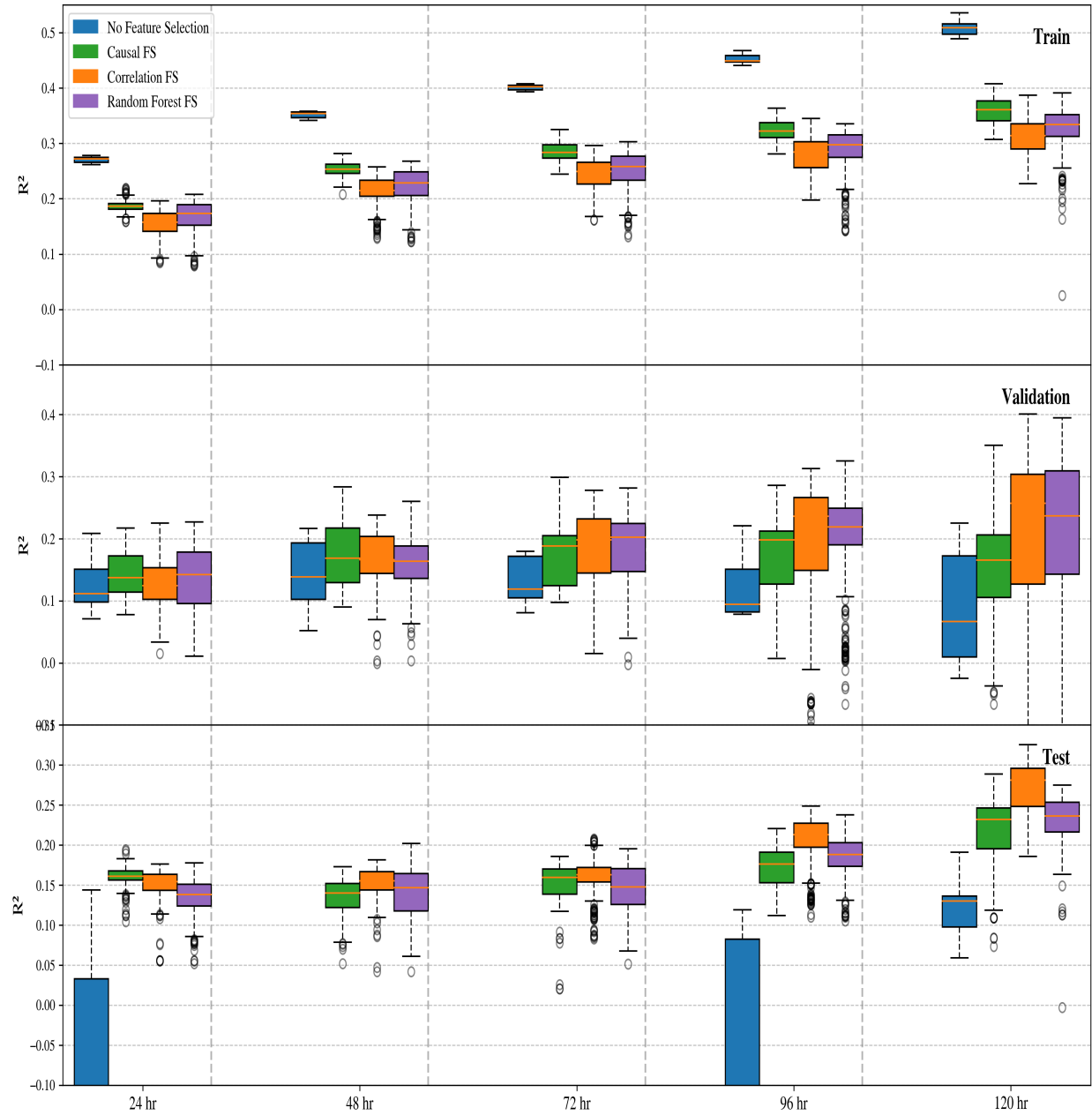
Figure S13: Boxplot comparing Train (Top), Validation (Middle) and Test (Bottom) R² values for DELV for each lead times 24, 48, 72, 96, 120 hrs for experiments without SHIPS link assumptions similar to Fig.3b, across four feature selection strategies: causal discovery, correlation ranking, random forest importance, and no selection. Causal feature selection yields the highest median R² till 72 hrs lead time, showing improved generalization in a purely statistical prediction setup.
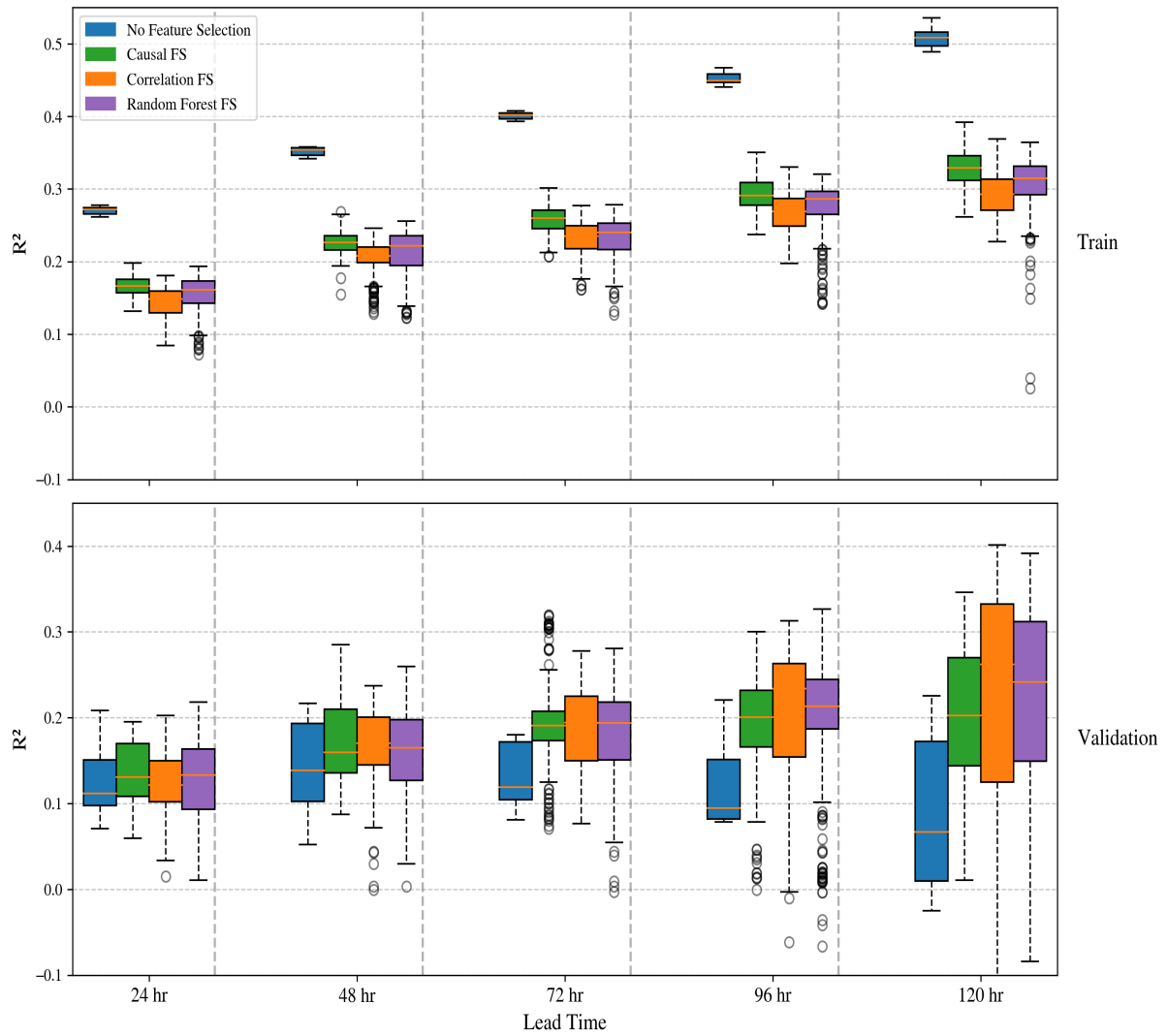
Figure S14: Boxplot comparing Training (Top), Validation (Bottom) R² values for DELV for each lead times 24, 48, 72, 96, 120 hrs for experiments with SHIPS link assumptions across four feature selection strategies: causal discovery, correlation ranking, random forest importance, and no selection. Causal feature selection yields the highest median R² till 72 hrs lead time, showing improved generalization in a purely statistical prediction setup.
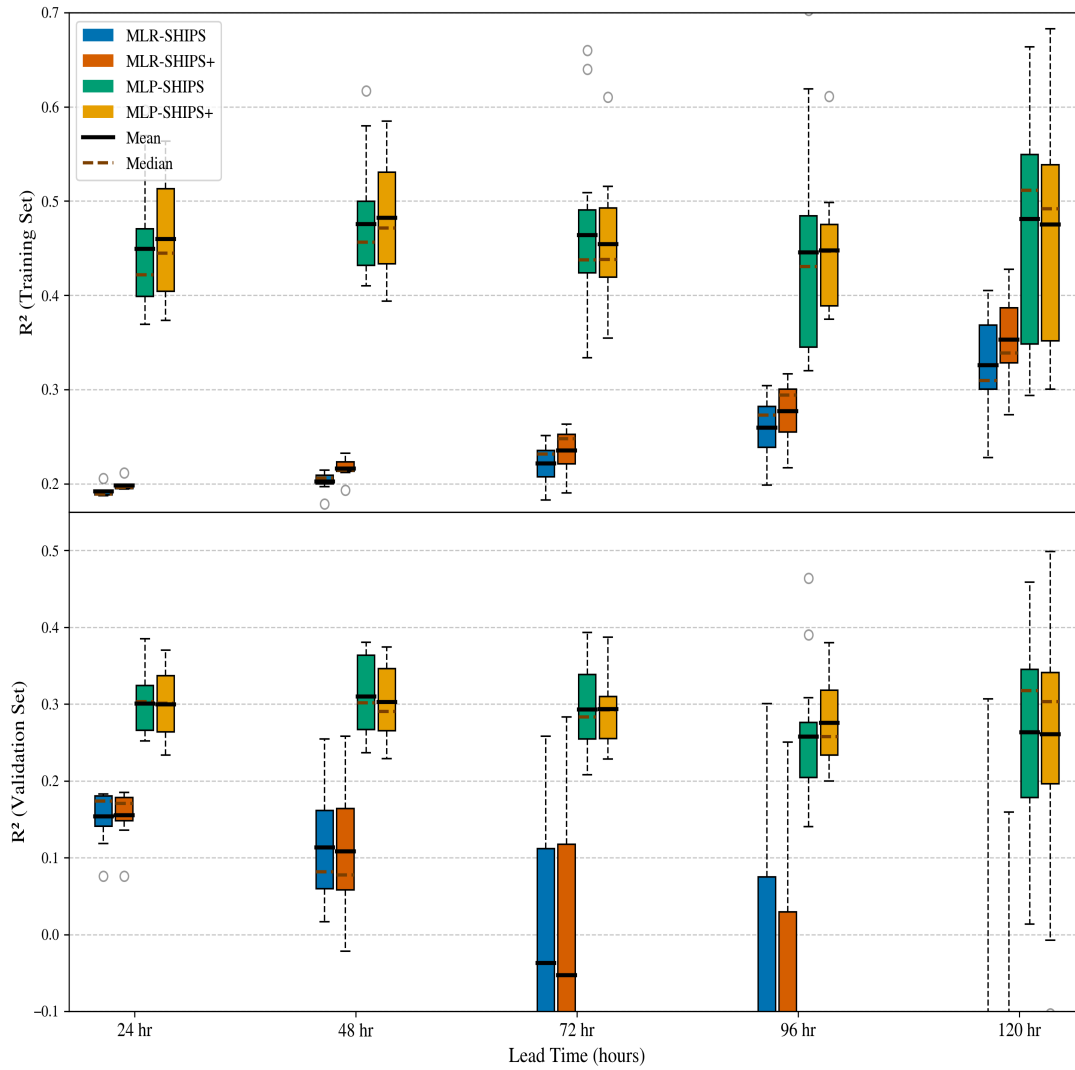
Figure S15: Boxplot comparing Train (Top), Validation (Bottom) R² values for DELV for each lead times 24, 48, 72, 96, 120 hrs for experiments using SHIPS and SHIPS+ predictors for MLR and MLP. MLP consistently outperforms and have the highest R² values for all lead times showing the nonlinear model's superior ability to capture complex relationships between predictors and intensity change.