A FEATURE-BASED EXPLANATION FRAMEWORK FOR TIME-DEPENDENT OUTPUTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Feature-based explanations quantify features' influence on model predictions, but are primarily designed for scalar outputs. In many applications, however, outputs are functional or multivariate, such as time-dependent trajectories in demand forecasting. Consequently, existing approaches typically explain each output location independently, ignoring dependencies across the output components. We address this limitation by developing a unified framework for feature-based explanations of time-dependent outputs. Specifically, we generalize functional decomposition to Hilbert-valued prediction functions and extend an existing feature-based explanation framework to this setting. Our framework introduces kernel-based output representations that enable time-dependency-aware explanations at multiple levels of temporal granularity, including *time-specific*, *time-resolved*, and *time-aggregated*, while providing a unified view in which existing methods arise as special cases.

1 INTRODUCTION

Machine learning models achieve strong predictive performance across many applications, but are widely considered black boxes. However, in high-stakes settings such as healthcare, finance, or policy making, understanding how predictions are formed is crucial (Rudin, 2019; Bracke et al., 2019; Tjoa & Guan, 2021). This has motivated explainable artificial intelligence (XAI), including feature-based methods that quantify how input features influence model predictions, either *locally* (individual predictions) (Lundberg & Lee, 2017; Goldstein et al., 2015) or *globally* (model behavior across the data distribution) (Breiman, 2001; Fisher et al., 2019; Covert et al., 2020). Recent work has sought to unify these methods and to reveal connections between them (Covert et al., 2021; Lundstrom & Razaviyayn, 2023; Deng et al., 2024; Fumagalli et al., 2025). In particular, the framework proposed by Fumagalli et al. (2025) and Herbinger et al. (2026) formalizes feature-based explanations through a common set of operators, providing a general view on existing approaches.

However, these methods and their unifying formulations are developed for *scalar-valued outputs*, as encountered in regression or classification tasks. In many applications, models instead produce *multivariate* or *functional outputs*, such as time-dependent trajectories (Salinas et al., 2020; Choi et al., 2016). In practice, XAI methods are routinely applied pointwise to functional outputs, explaining each output location in isolation (van Zyl et al., 2024; Neubauer et al., 2025; Gutierrez et al., 2024; Donizy et al., 2022). This has two limitations: it ignores *dependencies across the output domain*, treating each output location as if it were unrelated to its neighbors, and it offers no principled mechanism for *aggregating* feature effects *across time*. The result is a gap between the structure of time-dependent outputs and the levels of temporal granularity at which they can be explained.

To illustrate these limitations, consider a patient in intensive care where a biomarker (e.g., blood lactate) is monitored over time. A model F predicts a 24h trajectory (i.e., $F(\mathbf{x})$ is a function over t) (Fig. 1, top right), where X_1 controls the recovery trend, and X_2, X_3 represent clinically meaningful, early shock ($t = 10$h) and late deterioration ($t = 18$h) events, respectively. A natural goal is to understand how each feature influences the trajectory. Pointwise explanations attribute feature effects independently at each t. Yet, the influence of X_2 is not confined to $t = 10$h, but spreads over a temporal neighborhood, and clinicians may care less about attribution at a single moment than about the aggregate influence over a clinically meaningful phase. Pointwise methods support neither: they ignore dependencies across the output domain and lack principled basis for aggregation.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

To address these limitations, we propose the *Hilbert-valued explanation framework*, which extends feature-based explanations to multivariate and functional outcomes by encoding their dependence structure through a choice of kernel. Fig. 1 illustrates this contrast: an identity kernel produces isolated pointwise attributions, while a time-dependency-aware kernel (correlation kernel) yields phase-aware effects at a specific time, across the trajectory, or aggregated into a single importance.

Contributions.

1. Hilbert-valued functional decomposition: We introduce the Hilbert-valued functional decomposition ($\mathcal{H}$-FD), extending classical functional decomposition to time-dependent (multivariate or functional) outputs and yielding pure feature effects in a Hilbert space.

2. Hilbert-valued explanation framework: We extend the feature-based explanation framework of Fumagalli et al. (2025) to time-dependent outputs, capturing dependencies across the output domain and different levels of aggregation. Classical pointwise explanations and time-dependent Sobol indices arise as special cases, which serve as experimental baselines.

3. Practical guidance: We provide guidance on selecting appropriate kernels based on the desired interpretation and demonstrate their impact on the resulting explanations in real-world applications.

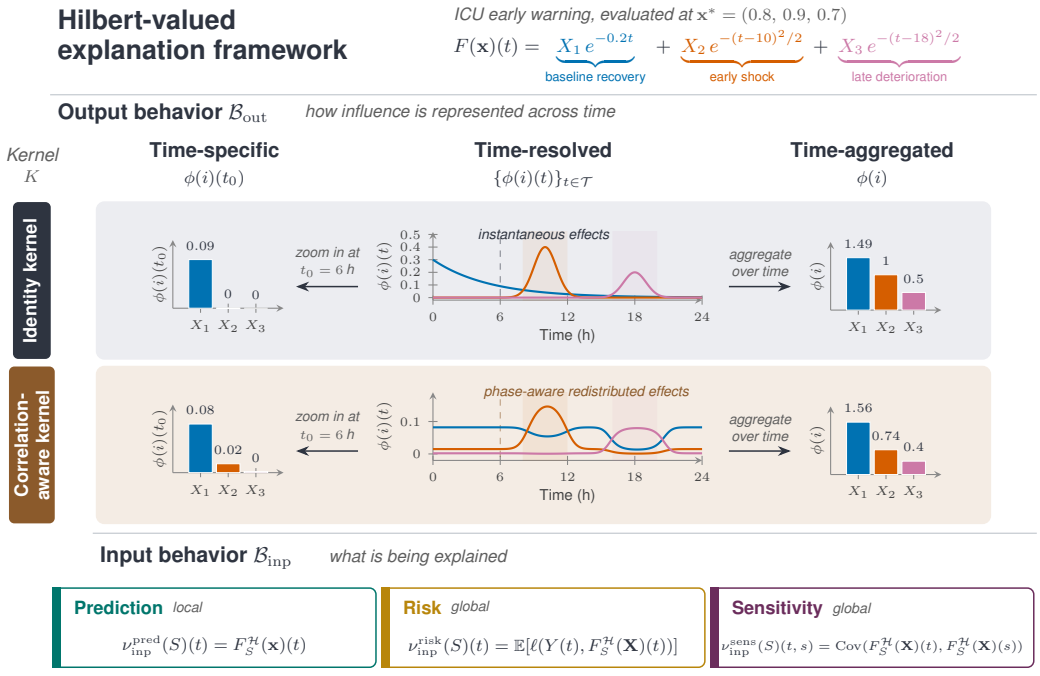


Figure 1: **Hilbert-valued explanation framework.** Illustration on an ICU early-warning toy model (*top*). Two orthogonal axes structure time-dependent attribution. The *output behavior* (middle) controls *how* influence is represented across time, parameterized by a kernel K: the identity kernel produces instantaneous effects, the pointwise baseline $\phi(i)(t)$ (*middle, row 1*), a time-dependency-aware kernel (correlation kernel) redistributes them phase-awarely (*middle, row 2*). *Resolved* trajectories can be sampled at *specific* time points t_0 or *aggregated* into a scalar $\phi(i)$. The *input behavior* (*bottom*) specifies *what* is being explained: prediction, risk, or sensitivity.

2 Related Work

Feature-based explanations for time-dependent outputs largely adapt prediction-level methods to forecasting settings by treating multi-step outputs independently. TsSHAP (Raykar et al., 2023) explains individual forecast time points, ShapTime (Zhang et al., 2023) focuses on aggregated quantities, and PAX-TS (Kreuzer et al., 2025) combines both perspectives. Beyond forecasting, SurvSHAP(t) (Krzyziński et al., 2023) explains time-dependent survival functions pointwise, with

SurvSHAP-IQ (Langbein et al., 2026) extending this to interaction effects. These methods remain confined to Shapley-style, prediction-level explanations operating on individual output locations or predefined aggregations, lacking a unified treatment of dependencies across the output domain.

Outside the XAI literature, sensitivity analysis has long addressed multivariate and functional outputs. Extensions of *Sobol indices* define importance measures by aggregating variance across the output domain (Campbell et al., 2006; Lamboni et al., 2011; Gamboa et al., 2014), while also accounting for temporal dependence through covariance-based formulations (Shi et al., 2018; Alexanderian et al., 2020). However, these approaches are restricted to variance-based notions of importance and have not been introduced from a feature-based XAI perspective.

3 BACKGROUND

Work on time-dependent outputs mainly arises in sensitivity analysis with limited links to feature-based explanations; we introduce the necessary concepts from both areas.

General notation. Consider a supervised learning setting with a function $F^{\mathcal{H}} : \mathcal{X} \to \mathcal{H}$, where $\mathcal{X} \subseteq \mathbb{R}^p$ and $\mathcal{H}$ is a Hilbert space with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. This covers *multivariate outputs* $F^{\mathcal{H}}(\mathbf{x}) \in \mathbb{R}^T$ ($T \in \mathbb{N}$) and *functional outputs*, $F^{\mathcal{H}}(\mathbf{x}) \in L^2(\mathcal{T})$ with $\mathcal{T} \subseteq \mathbb{R}$. Standard scalar-valued outputs are written $F : \mathcal{X} \to \mathbb{R}$. For $t \in \mathcal{T}$, define $F_t(\mathbf{x}) := F^{\mathcal{H}}(\mathbf{x})(t)$. Let $\mathbf{X} = (X_1, \ldots, X_p)$ denote the input random vector. For a subset $S \subseteq \{1, \ldots, p\}$, let $\mathbf{X}_S$ denote subvectors and by $\mathbf{X}_{-S}$ complements. While the theory applies to general Hilbert-valued outputs, including both finite- and infinite-dimensional settings, we use time-dependent trajectories as the running case throughout.

3.1 SENSITIVITY ANALYSIS

Sensitivity analysis attributes output variability to input features and their interactions via functional decomposition, inducing a variance decomposition that underlies closed and total Sobol indices.

Functional and variance decomposition. Let $F : \mathcal{X} \to \mathbb{R}$ denote a prediction function and let $P_{\mathbf{X}}$ be a reference distribution over $\mathcal{X}$. Functional decomposition represents F as a sum of pure effects f_S over all feature subsets $S \subseteq \{1, \ldots, p\}$, where each f_S captures effects beyond its subsets,

$$F(\mathbf{x}) = \sum_{S \subseteq \{1,\ldots,p\}} f_S(\mathbf{x}), \quad \text{with} \quad f_S(\mathbf{x}) := F_S(\mathbf{x}) - \sum_{L \subset S} f_L(\mathbf{x}) \tag{1}$$

and $F_S(\mathbf{x}) = \mathbb{E}_{\mathbf{X}_{-S} \sim P_{\mathbf{X}_{-S}}}[F(\mathbf{X}) \mid \mathbf{X}_S = \mathbf{x}_S]$ denotes the marginal effect of features in S (Hoeffding, 1948; Sobol', 2001; Hooker, 2004). The decomposition depends on a reference distribution, typically marginal or conditional (Fumagalli et al., 2025). Under feature independence and square-integrability of F, the components are orthogonal and induce the variance decomposition $\mathbb{V}[F(\mathbf{X})] = \sum_{S \subseteq \{1,\ldots,p\}} \mathbb{V}[f_S(\mathbf{X})]$. For time-dependent outputs, this decomposition applies pointwise via $F_t(\mathbf{x}) = F^{\mathcal{H}}(\mathbf{x})(t)$ with f_L replaced by $f_{L,t}(\mathbf{x}) = f_L^{\mathcal{H}}(\mathbf{x})(t)$.

Based on this variance decomposition, Sobol indices quantify the contribution of feature subsets to the total variance. The **closed Sobol index** for a subset S is defined as

$$\xi_S = \frac{\mathbb{V}[F_S(\mathbf{X})]}{\mathbb{V}[F(\mathbf{X})]} = \frac{\sum_{L \subseteq S} \mathbb{V}[f_L(\mathbf{X})]}{\mathbb{V}[F(\mathbf{X})]}, \tag{2}$$

where the second equality holds under independent inputs (Sobol', 1993; 2001). It captures all effects involving only features in S, while the **total Sobol index** additionally includes all interaction effects involving at least one feature in S (Saltelli et al., 2008).

Time-dependent Sobol indices. For multivariate or functional outputs, variance is replaced by the covariance kernel $\mathrm{Cov}(F_t, F_s)$, leading to Sobol-type indices that account for dependencies across the output domain. A **time-specific** *closed* Sobol index can then be defined as (Shi et al., 2018):

$$\xi_S^{\mathrm{spec}}(t) = \frac{\int_{\mathcal{T}} \mathrm{Cov}\big(F_{S,t}(\mathbf{X}), F_{S,s}(\mathbf{X})\big)\, ds}{\int_{\mathcal{T}} \mathrm{Cov}\big(F_t(\mathbf{X}), F_s(\mathbf{X})\big)\, ds}, \tag{3}$$

accounting for time-dependencies relative to a fixed t. The corresponding *total* index is obtained analogously by aggregating all effects involving at least one feature in S. Integrating these covariance contributions over time, yields scalar **time-aggregated** closed and total Sobol indices (Alexanderian et al., 2020). Neglecting cross-time covariance reduces it to a sum of marginal variances.

3.2 Feature-based Explanations

Feature-based explanations quantify how features influence model behavior. Following Fumagalli et al. (2025); Herbinger et al. (2026), they can be expressed as compositions of *masking* ($\mathcal{M}$), *behavior* ($\mathcal{B}$), and *interaction* ($\mathcal{I}$) operators.

Masking operator. The masking operator $\mathcal{M}$ constructs a predictor depending only on features in S,

$$F_S(\mathbf{x}) := (\mathcal{M}F)(\mathbf{x}, S),$$

by removing (masking) features in $-S$ using a reference distribution $P_{X_{-S}}$ (e.g., marginal or conditional). Notably, F_S corresponds to the marginal effect used in Eq. 1.

Behavior operator. The behavior operator $\mathcal{B}$ maps masked predictors to a set function ν

$$\nu(S) := (\mathcal{B}(\mathcal{M}F))(S),$$

which defines the quantity of interest, yielding local predictions $F_S(\mathbf{x})$ or global quantities such as prediction variance $\mathrm{Var}(F_S(\mathbf{X}))$ or expected loss $\mathbb{E}[\ell(Y, F_S(\mathbf{X}))]$.

Interaction operator. The interaction operator $\mathcal{I}$ maps ν to feature-based explanations ϕ

$$\phi(j) := (\mathcal{I}(\mathcal{B}(\mathcal{M}F)))(j), \qquad j \in \{1, \dots, p\}.$$

This operator quantifies individual feature influence while accounting for higher-order effects encoded in ν. This yields different notions of explanations: *pure effects*, $\phi(j) = \nu(j) - \nu(\emptyset)$, quantify the standalone contribution of feature j, excluding interactions; *partial effects*, $\phi(j) = \frac{1}{p}\sum_{S \subseteq -j} \binom{p-1}{|S|}^{-1} [\nu(S \cup j) - \nu(S)]$, correspond to Shapley values and distribute interactions across features; and *full effects*, $\phi(j) = \nu(\{1, \dots, p\}) - \nu(-j)$, capture the total contribution of j, including interactions. Extensions to feature group explanations are defined in Fumagalli et al. (2025).

This operator view recovers many established methods as special cases. Prediction-based behavior yields Shapley values with partial effects (Shapley, 1953) and (conditional) partial dependence (PD) with pure effects (Friedman, 2001), while risk-based behavior with full effects yields (conditional) permutation feature importance (PFI) (Fisher et al., 2019; Strobl et al., 2008). Closed and total Sobol indices similarly arise from marginal masking, variance-based behavior, and pure or full effects, linking to classical sensitivity analysis. An overview is provided in Tab. 2, with more details in Fumagalli et al. (2025).

Yet, this framework is limited to scalar outputs F, leaving Hilbert-valued outputs $F^{\mathcal{H}}$ unaddressed.

4 Methodology

We extend feature-based explanations to Hilbert-valued outputs via a functional decomposition in $\mathcal{H}$ to provide a general theoretical framework that covers a broad range of time-dependent output types.

4.1 Hilbert-valued Functional Decomposition ($\mathcal{H}$-FD)

We generalize the functional decomposition of Eq. 1 to Hilbert-valued outputs $F^{\mathcal{H}}(\mathbf{x})$. This formulation treats finite-dimensional multivariate outputs, $\mathcal{H} = \mathbb{R}^T$, and functional outputs, such as $\mathcal{H} = L^2(\mathcal{T})$, within the same framework; in particular, the construction does not depend on a fixed discretization of the output domain. The decomposition is performed over the input space as before, while the resulting effects take values in a Hilbert space $\mathcal{H}$. Intuitively, each effect contributes not a single value but a *function* (or vector) describing its influence across the output domain.

Definition 1 (Hilbert-Valued Functional Decomposition)**.** *The Hilbert-valued prediction function $F^{\mathcal{H}}(\mathbf{x})$ admits a decomposition into **Hilbert-valued pure effects** $f_S^{\mathcal{H}}$:*

$$F^{\mathcal{H}}(\mathbf{x}) = \sum_{S \subseteq \{1,\dots,p\}} f_S^{\mathcal{H}}(\mathbf{x}), \quad f_S^{\mathcal{H}}(\mathbf{x}) := F_S^{\mathcal{H}}(\mathbf{x}_S) - \sum_{L \subset S} f_L^{\mathcal{H}}(\mathbf{x}). \tag{4}$$

Here, $F_S^{\mathcal{H}}$ denotes the Hilbert-valued marginal effect with respect to a reference distribution $P_{\mathbf{X}_{-S}}$:

$$F_S^{\mathcal{H}}(\mathbf{x}_S) := \mathbb{E}_{\mathbf{X}_{-S} \sim P_{\mathbf{X}_{-S}}}[F^{\mathcal{H}}(\mathbf{X}) \mid \mathbf{X}_S = \mathbf{x}_S], \quad \textit{where} \quad F_\emptyset^{\mathcal{H}} := \mathbb{E}[F^{\mathcal{H}}(\mathbf{X})] \tag{5}$$

Thus, $\mathcal{H}$-FD follows the scalar construction, but the effects $f_S^{\mathcal{H}}(\mathbf{x})$ are elements (e.g., vectors or functions) in $\mathcal{H}$.

4.2 HILBERT-VALUED EXPLANATION FRAMEWORK

Now, we extend the framework introduced in Sec. 3.2 to Hilbert-valued outputs $F^{\mathcal{H}}$ based on the $\mathcal{H}$-FD. A detailed overview of operator choices is provided in App. B.

The **masking operator** $\mathcal{M}$ remains unchanged and connects directly to the $\mathcal{H}$-FD: the masked predictor $F_S^{\mathcal{H}}(\mathbf{x}) := (\mathcal{M}F^{\mathcal{H}})(\mathbf{x}, S)$ corresponds to the marginal effect and is itself a function (or vector) over the output domain.

Since $F_S^{\mathcal{H}}$ is defined over the output domain, the **behavior operator** $\mathcal{B}$ must specify both *what* is explained and *how* it is represented across that domain, in particular with respect to *dependencies between output locations* and the *level of aggregation*. We therefore decompose it as

$$\mathcal{B}(F_S^{\mathcal{H}}) := \mathcal{B}_{\text{out}}\big(\mathcal{B}_{\text{inp}}(F_S^{\mathcal{H}})\big) = (\mathcal{B}_{\text{out}} \circ \mathcal{B}_{\text{inp}})(F_S^{\mathcal{H}}),$$

where $\mathcal{B}_{\text{inp}}$ defines the quantity of interest in the **input space** (as in the scalar case), and $\mathcal{B}_{\text{out}}$ specifies its representation over the **output domain**.

This yields pipeline extended from Fumagalli et al. (2025); Herbinger et al. (2026)

$$F^{\mathcal{H}} \xrightarrow{\mathcal{M}} F_S^{\mathcal{H}} \xrightarrow{\mathcal{B}_{\text{inp}}} \nu_{\text{inp}}(S) \xrightarrow{\mathcal{B}_{\text{out}}} \nu(S) \xrightarrow{\mathcal{I}} \phi(j).$$

Input behavior operator $\mathcal{B}_{\text{inp}}$. The input behavior operator maps masked predictors to Hilbert-valued quantities of interest,

$$\nu_{\text{inp}}(S) := (\mathcal{B}_{\text{inp}}(\mathcal{M}F^{\mathcal{H}}))(S),$$

analogous to the scalar case. Typical choices include local prediction-based behavior $\nu_{\text{inp}}^{\text{pred}}(S)(t) = F_S^{\mathcal{H}}(\mathbf{x})(t)$, global risk-based behavior $\nu_{\text{inp}}^{\text{risk}}(S)(t) = \mathbb{E}[\ell(Y(t), F_S^{\mathcal{H}}(\mathbf{X})(t))]$, and sensitivity-based behavior $\nu_{\text{inp}}^{\text{sens}}(S)(t,s) = \text{Cov}(F_S^{\mathcal{H}}(\mathbf{X})(t), F_S^{\mathcal{H}}(\mathbf{X})(s))$. Depending on the choice, $\nu_{\text{inp}}(S)$ is a function over $\mathcal{T}$ or a covariance surface over $\mathcal{T} \times \mathcal{T}$.

Output behavior operator $\mathcal{B}_{\text{out}}$. The output behavior operator determines how Hilbert-valued quantities are represented and aggregated across the output domain. The Hilbert-space structure provides an inner product for comparing output elements and supports kernel-induced operators that encode dependencies between output locations. We use a symmetric positive semi-definite kernel K, whose choice determines how dependencies across the output domain are incorporated into the explanation and thus depends on the desired interpretation. For $\mathcal{H} = \mathbb{R}^T$, the resulting kernel operation reduces to standard matrix multiplication, while the same construction applies directly to functional output spaces:

$$\nu(S) := (\mathcal{B}_{\text{out}}(\nu_{\text{inp}}))(S), \quad (\mathcal{B}_{\text{out}}^{\text{spec}}\nu_{\text{inp}})(S,t) := \int_{\mathcal{T}} K(t,s)\, \nu_{\text{inp}}(S)(\cdot)\, ds. \tag{6}$$

The time-specific operator $\mathcal{B}_{\text{out}}^{\text{spec}}$ defines the base construction, yielding a **time-specific** explanation at each t, which may incorporate information from other time points through K. Evaluating it for all $t \in \mathcal{T}$ gives **time-resolved** explanations, while aggregating over $\mathcal{T}$, i.e., $(\mathcal{B}_{\text{out}}^{\text{aggr}}\nu_{\text{inp}})(S) = \int_{\mathcal{T}} (\mathcal{B}_{\text{out}}^{\text{spec}}\nu_{\text{inp}})(S,t)\, dt$, yields **time-aggregated** explanations.

For time-aggregated explanations, kernel aggregation admits an explicit integral representation, characterizing when feature rankings are invariant to the choice of kernel.

Theorem 1. *Let $\mathcal{B}_{\text{out}}^{\text{spec}}$ be defined as in Eq. 6. Then, for any subset $S \subseteq \{1, \dots, p\}$, the time-aggregated effect with prediction-based input behavior $\nu_{\text{inp}}(S)(t) = F_S^{\mathcal{H}}(\mathbf{x})(t)$ admits*

$$(\mathcal{B}_{\text{out}}^{\text{aggr}}\nu_{\text{inp}})(S) = \int_{\mathcal{T}} w_K(s)\, \nu_{\text{inp}}(S)(s)\, ds, \qquad w_K(s) := \int_{\mathcal{T}} K(t,s)\, dt.$$

*If $\mathcal{B}_{\text{inp}}$ is **linear** in $F^{\mathcal{H}}$ and $\nu_{\text{inp}}(\{i\})(s) \geq \nu_{\text{inp}}(\{j\})(s)$ for all $s \in \mathcal{T}$, the corresponding time-aggregated pure effects **preserve the ordering** of features i and j for any kernel K satisfying $K(t,s) \geq 0$. Analogous results hold for partial and full effects; see App. A.1.*

Remark 1. *For **nonlinear** input behaviors such as sensitivity or risk, the ordering invariance in Thm. 1 does not generally hold. Although the same weighted-integral representation applies, different kernels may induce different feature rankings.*

Finally, note that the proposed framework preserves the additive decomposition: combining masking with input and output behavior operators always yields a valid set function over feature subsets, ensuring that feature-based explanations remain well-defined.

Proposition 1. *Let* $\nu(S) := (\mathcal{B}_{\text{out}}(\mathcal{B}_{\text{inp}}(\mathcal{M}F^{\mathcal{H}})))(S)$. *Then the Möbius decomposition holds:*

$$\nu(\{1,\dots,p\}) = \sum_{S\subseteq\{1,\dots,p\}} \sum_{L\subseteq S} (-1)^{|S|-|L|}\nu(L)$$

(Rota, 1964; Grabisch & Roubens, 1999). Thus, time-specific, time-resolved, and time-aggregated explanations admit a unique additive decomposition for fixed choices of $\mathcal{M}$, $\mathcal{B}_{\text{inp}}$, *and* $\mathcal{B}_{\text{out}}$.

Interaction operator and resulting explanations. The interaction operator $\mathcal{I}$ maps the set function ν to feature-based explanations $\phi(j) := (\mathcal{I}(\nu))(j)$. As illustrated in Fig. 1, the resulting explanation depends on both behavior operators: $\mathcal{B}_{\text{inp}}$ determines *what* is explained, while $\mathcal{B}_{\text{out}}$ determines *how* feature influence is represented across time. This yields scalar-valued time-specific and time-aggregated explanations, as well as time-resolved explanations as functions (or vectors) over $\mathcal{T}$, with the choice of kernel determining how temporal dependencies are incorporated.

4.3 Unifying Existing Methods

Time point feature-based explanations. If $K = I$ (identity kernel), the output behavior $\mathcal{B}_{\text{out}}$ ignores cross-time dependencies: time-specific and time-resolved explanations reduce to pointwise explanations, and time-aggregated explanations to simple averages over the output domain. Within this setting, existing methods arise as special cases combining $K = I$ with prediction-based input behavior (see App. C). In particular, TsSHAP corresponds to time-specific Shapley values. ShapTime operates on aggregated output quantities, yielding time-aggregated explanations, while PAX-TS provides time-specific and time-aggregated explanations. SurvSHAP(t) produces time-resolved explanations over the trajectory; SurvSHAP-IQ extends this to interaction effects via the operator $\mathcal{I}$. We use pointwise explanations, including TsSHAP and ShapTime, as baselines in Sec. 5.2.

In summary, existing approaches are confined to local prediction-based explanations and are either pointwise (time-specific/time-resolved) or based on predefined aggregations (time-aggregated), without explicitly modeling dependencies across the output domain. Our framework removes both restrictions: risk- and sensitivity-based input behaviors yield global notions of importance, while general kernels K enable time-dependency-aware explanations.

Time-dependent Sobol indices. Beyond prediction-based methods, our framework also recovers Sobol measures for Hilbert-valued outputs via sensitivity-based input behavior and a constant kernel.

Theorem 2. *Let* $\nu_{\text{inp}}(S)(t,s) = \text{Cov}\big(F_S^{\mathcal{H}}(\mathbf{X})(t), F_S^{\mathcal{H}}(\mathbf{X})(s)\big)$ *and choose a constant kernel* $K(t,s) = 1$. *Then the output behavior yields*

$$\nu(S)(t) = \int_{\mathcal{T}} \nu_{\text{inp}}(S)(t,s)\,ds, \qquad \nu(S) = \int_{\mathcal{T}}\int_{\mathcal{T}} \nu_{\text{inp}}(S)(t,s)\,dt\,ds.$$

Under feature independence, the corresponding pure and full interaction operators $\mathcal{I}$ *recover the closed and total time-specific and time-aggregated Sobol indices.*

5 Experiments

We evaluate the proposed framework on synthetic and real-world datasets, with a *focus on kernel choice* and its impact on resulting explanations. Implementations are available at GitHub, with full computational details and empirical validation of *feature effects and Sobol indices recovery* (Thm. 2) against known ground truth across model classes and sample sizes in App. D.

5.1 Kernel Guidance

The kernel K encodes assumptions about how feature influence is distributed over the output domain; it should reflect the *research question* rather than be tuned as a hyperparameter. We summarize guidance via two principles, illustrated on synthetic scenarios using pure first-order local effects $f_i^{\mathcal{H}}(\mathbf{x})$ for a random observation (Fig. 2; see also App. D.2, Tab. 5 for an extended kernel guidance).

Principle 1: Match temporal scope to the notion of contribution. If the question is *which feature drives the output at an instant*, use the identity kernel ($K = I$). For *sustained influence* or *alignment with trajectory shape*, use smoothing or correlation-aware kernels. In the ICU example (Fig. 2, row 1), the identity kernel yields sharp peaks for X_2 and X_3 at $t = 10$h and $t = 18$h, while the OU kernel $K_{\mathrm{OU}}(t, s) = \sigma^2 \exp(-|t - s|/\ell)$ broadens these into windows of width ℓ, reflecting effects that persist around a shock. The correlation kernel weights effects by alignment with output covariance, amplifying X_1 during its slow exponential decay, which drives the trajectory between shocks, while retaining localized peaks for X_2 and X_3. These are not competing estimates — they answer *different questions*, each valid in context.

Principle 2: Match kernel symmetry to temporal causality. Symmetric kernels (Gaussian, OU) implicitly assume that influence at t can be informed by past and future output values, which can be temporally incoherent for causal processes. Fig. 2, row 2 considers a market price response driven by an unprecedented and short-lived pulse X_1 active on $[0.5, 1.0)$h, but a Gaussian kernel assigns attribution before the pulse occurs. A causal exponential kernel $K_{\mathrm{causal}}(t, s) = \exp(-(t-s)/\ell)\, \mathbf{1}_{t \geq s}$ restricts attribution to $t \geq 0.5$h. In general, kernel structure should mirror the data-generating process: symmetric for bidirectional dependence, causal for events, interventions, or autoregressive dynamics. When features differ in temporal structure, a single kernel imposes the same assumption on all of them, and kernels should be selected per feature, as in the SPY application in Sec. 5.2.

When kernel choice matters less. Under prediction input behavior, time-aggregated rankings are stable across kernels whenever attribution trajectories maintain a consistent ordering over time (Thm. 1), as in both examples; the kernel then affects only the temporal shape. This does not extend to sensitivity- or risk-based behavior, where kernel choice can reorder features (App. D.1, Fig. 8).

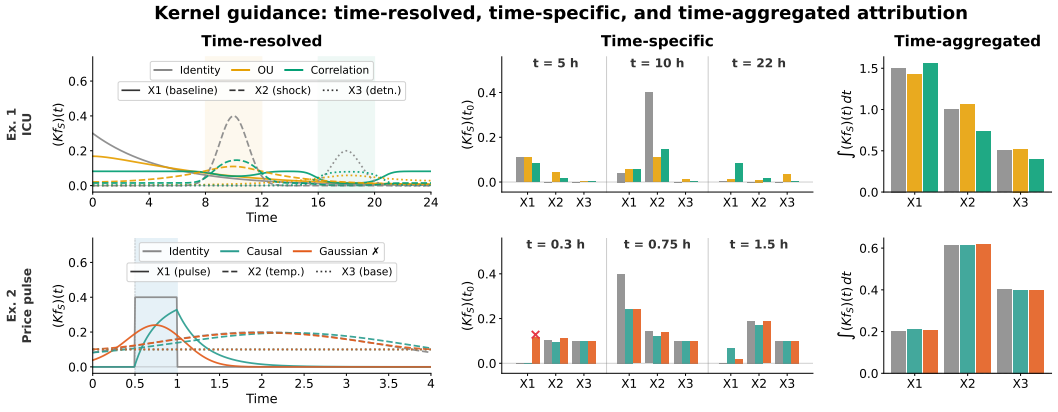


Figure 2: Kernel guidance for two synthetic examples with distinct temporal structures: an ICU early warning model (*row 1*) and a market price pulse (*row 2*). Each shows *pure local effects* at three explanation levels: time-resolved (*col. 1*), time-specific at selected times (*col. 2*), and time-aggregated (*col. 3*) for marginal masking $\mathcal{M}$, prediction-based input behavior $\mathcal{B}_{\mathrm{inp}}$, and pure interaction operator $\mathcal{I}$ of a randomly chosen observation; identity-kernel curves are the pointwise baselines underlying existing methods (here, pointwise centered PDPs; see Sec. 3.2 and App. C).

5.2 Real-world Examples and Baseline Comparison

Intraday SPY volatility prediction. We predict intraday volatility for the SPDR S&P 500 ETF (SPY), using five-minute absolute log-returns (78 time steps) (a standard volatility proxy) and six pre-market features, including prior-day VIX [1] (`vix_prev`) and a macro announcement indicator (`ann_indicator`). A multivariate random forest is trained on 560 trading days (Jan 2022–Apr 2024) (Polygon.io, 2024). We subtract the average diurnal pattern to focus on deviations from the typical U-shape. Fig. 3 examines July 13, 2022, a high-volatility day during the Fed's rate-hiking cycle, when the Beige Book — a qualitative indicator of US economic conditions – was released at 14:00 ET; full results are in App. D.3. For a trader or risk manager, the key question is not just *which* features matter, but *when* and *how* they drive volatility over the day.

[1] CBOE Volatility Index (VIX), the so-called "fear index".

Time-aggregated pure and partial local effects (Fig. 3, col. 1) identify `vix_prev` as the dominant driver; thus the ShapTime baseline (row 1, col. 1, partial) agree with the kernel-weighted aggregated effects (row 2, col. 1, partial). We additionally analyze `ann_indicator` given the scheduled Beige Book release. The temporal structure of their effects depends critically on the kernel. Under the identity kernel (row 1), corresponding to existing methods or their pointwise counterparts (Sec. 4.3; App. C), pure effects (col. 2), i.e. pointwise PD effects, are noisy and difficult to interpret. Temporally structured kernels (row 2) reveal economically meaningful patterns: For `vix_prev`, an OU kernel yields a smooth U-shape, indicating elevated volatility at the open and close. In contrast, a causal kernel localizes `ann_indicator` at 14:00 without pre-event leakage, whereas a symmetric smoother such as OU would induce spurious pre-announcement attribution. The partial pointwise baseline (i.e., TsSHAP in row 1, col 3) also recovers the spike but assigns no post-event influence, discarding the economically meaningful decay. App. D.5 quantifies this across 19 intraday-release days: the causal kernel consistently places less attribution mass before the release than the pointwise baseline or matched symmetric kernel, providing a quantitative measure of temporal localization. Partial effects (col. 3) show that the time-aggregated contribution of `ann_indicator` is $1.61\times$ its pure effect, indicating strong interaction with other features. The interaction with `vix_prev` (col. 4) reveals a clear regime shift: before 14:00, the features are mildly redundant, both reflecting elevated expected volatility; afterward, they become strongly synergistic, with the release amplifying the impact of prior market fear. A scalar interaction measure would average over this shift and miss the regime change. Overall, the framework reveals when volatility drivers act and how their role shifts intraday, matching traders' interpretation of market events.

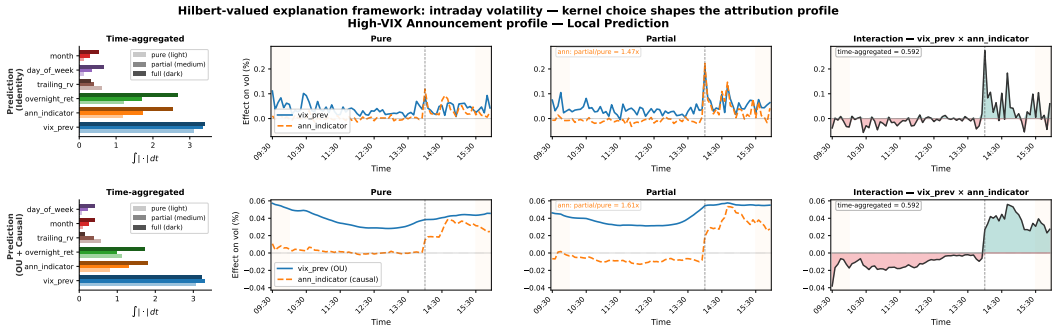


Figure 3: Local prediction effects for July 13, 2022 (`vix_prev` $= 27.3$, `ann_indicator` $= 1$), showing `vix_prev` and `ann_indicator`. *Col. 1*: Time-aggregated pure, partial, and full importances across all features under the identity kernel (*row 1*) and OU/causal kernel (*row 2*). *Col. 2–4*: Pure, partial effects and pure interaction effects, respectively. The dashed vertical line marks the 14:00 ET Beige Book release. Pure prediction-based input behavior recovers the PD effect and partial prediction-based input behavior recovers Shapley values (see App. B); thus, *row 1, col. 1* corresponds to ShapTime, *row 1, col. 2* to pointwise PD effects and *row 1, col. 3* to TsSHAP.

Electricity demand comparison. Grid operators must manage forecast uncertainty, as high demand variance requires costly reserves. We ask which features drive this uncertainty and whether it is localized or spans day. This can be studied using the full sensitivity effect (time-dependent total Sobol index, see Thm. 2). The kernel determines how prediction variance is attributed over time: the identity kernel treats each time step independently yielding pointwise (time-specific) Sobol indices, while the correlation kernel captures temporal dependencies. We evaluate on the UCI Individual Household Electric Power Consumption dataset (IHEPC, $T = 24$ hourly periods, $\approx$1400 days, one *individual* household) (Dua & Graff, 2019) and the GB National Electricity System Operator historic demand dataset (NESO, $T = 48$ half-hourly periods, $\approx$1800 days, *national* demand) (National Energy System Operator, 2024), following the SPY modeling setup (full details in App. D.4).

The datasets differ markedly in temporal dependence (Fig. 4, col. 1 and 4). IHEPC shows strong correlation between nearby time points decaying with temporal distance, while NESO demand is strongly correlated across the entire day. The network plots (col. 2 and 5) summarize full sensitivity effects, with node size encoding total Sobol effects and edge width full pairwise interaction strength. IHEPC is most strongly driven by `lag_daily_mean` (`LDM`), with diffuse interactions reflecting the idiosyncratic nature of single-household demand. NESO variance is dominated by `month`

(Mon) and season (Sea), with stronger interactions with each other, lag_evening (LEv) and lag_morning (LMo). The two kernels then tell complementary stories. For IHEPC, identity kernel full sensitivity effects reveal sharp peaks in the AM and PM high-demand windows (red and blue shaded areas), whereas the correlation kernel smooths these into a shallower, broader elevation, since nearby hours are tightly coupled and attribution at any peak hour borrows strength from its neighbors (col. 3). For NESO, intraday variation under the identity kernel collapses to near-flat trajectories under the correlation kernel, consistent with a regime shift affecting the entire day (col. 6). In both cases, hourly reserve adjustments based on identity-kernel results mis-target the structure of forecast variance: IHEPC calls for peak-window provisioning, NESO for daily capacity planning. Overall, the correlation kernel matches effect resolution to the dataset's dependence scale: local smoothing for IHEPC, full-day coherence for NESO.

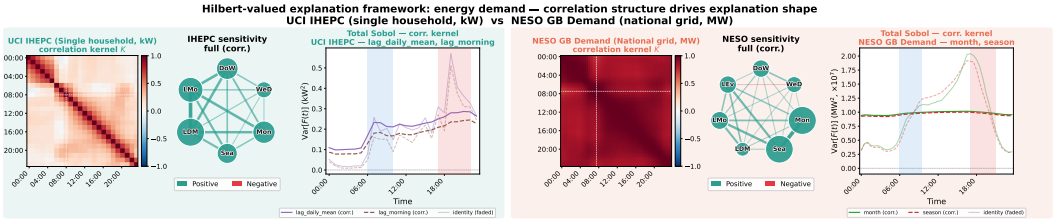


Figure 4: For each dataset (teal: IHEPC single household; orange: NESO national grid): empirical correlation heatmap (*col. 1*); full sensitivity network under the correlation kernel (*col. 2*); full sensitivity-based effect (total Sobol) for the two dominant features under the identity kernel (faded; pointwise Sobol baseline) and correlation kernel (solid/dashed) (*col. 3*). Blue and red shading marks AM and PM peak windows.

6 Discussion and Limitations

We introduced a Hilbert-valued explanation framework that generalizes feature-based explanations to time-dependent outputs, where the kernel choice directly affects the explanations obtained, surfacing temporal structure that pointwise approaches (e.g., TsSHAP or ShapTime) cannot capture. The kernel and its hyperparameters are not tuning parameters for predictive performance; they encode the notion of *contribution* — instantaneous, persistent, phase-dependent, or causal — and that should therefore be chosen based on the scientific question and domain knowledge (see Tab. 5).

This flexibility also introduces a limitation: a kernel misaligned with the intended interpretation may lead to misleading conclusions. Kernel choice therefore requires careful consideration and, where possible, domain-informed parameter selection. Computational cost is another consideration when using Shapley-based interaction operators (Lundberg & Lee, 2017; Covert et al., 2021), although standard approximation methods apply (Štrumbelj & Kononenko, 2014) and fast implementations can be extended (Jethani et al., 2022; Fumagalli et al., 2023). Pure and full effects, in contrast, can be computed more efficiently. More details on computational costs and experimental runtimes are provided in App. D.7 and App. D.9.

While our experiments instantiate the finite-dimensional case $\mathcal{H} = \mathbb{R}^T$, as is common for trajectories observed on a finite grid, the framework itself is formulated for general Hilbert-valued outputs and does not depend on a particular discretization. It therefore applies directly to genuinely functional outputs, such as continuous trajectories, spatial fields, spectra, or distribution-valued outputs, through an appropriate choice of output space and kernel. Exploring such applications empirically is an important direction for future work. Beyond this, a promising direction is to combine it with (functional) principal component analysis, to obtain low-dimensional and more interpretable representations of output variability before attribution. Another avenue is the use of operator-valued kernels (Kadri et al., 2016), allowing output dependencies to vary with the input. Finally, the framework can accommodate time-dependent inputs through suitable masking choices, which we leave for future work.

Broader Impact and Ethics Statement

This work introduces a methodological framework for explaining machine learning models with time-dependent outputs. As a post-hoc, model-agnostic explanation method, it does not itself produce predictions or decisions, but rather aids interpretation of existing models. We see two main avenues of positive impact. First, in high-stakes domains where models with functional outputs are increasingly deployed — clinical trajectory prediction, energy demand forecasting, financial risk modeling — our framework supports more transparent model auditing by making temporal dependencies in feature attributions explicit, which scalar pointwise methods obscure. Second, by unifying existing approaches under a common operator view, the framework lowers the barrier to comparing explanations across methods and may reduce ad-hoc method selection in applied work.

Limitations and risks of explanation methods generally apply here. Explanations can create unwarranted confidence in model behavior, particularly when the underlying model is itself unreliable or trained on biased data; our framework does not address these upstream issues and should not be treated as a substitute for model validation, fairness analysis, or domain expertise. The kernel choice introduced by our framework adds an additional modeling decision: a misspecified kernel can produce attributions that are technically valid but misleading for the question at hand (for example, applying a symmetric kernel to a causal process, as we illustrate in our synthetic experiments). We address this through explicit kernel-selection guidance (Tab. 5), but practitioners retain responsibility for justifying their choices. As with all Shapley-based methods, our explanations rely on a reference distribution and inherit known concerns about marginal versus conditional masking under feature dependence; users should select the masking strategy appropriate to their interpretive goal.

Our experiments use publicly available datasets (UCI IHEPC, GB NESO historic demand) and a standard licensed financial dataset (Polygon.io intraday equity bars). None involve human subjects, personally identifiable information, or sensitive demographic attributes. The IHEPC dataset contains household-level electricity consumption from a single anonymized household; we use it solely as a benchmark for the methodological framework and draw no inferences about individuals. We do not foresee direct dual-use concerns: the framework is a diagnostic tool for existing models rather than an enabling capability for new ones.

Reproducibility Statement

All assumptions underlying our theoretical claims are stated in Sec. 4; complete proofs of Thm. 1, Prop. 1, and Thm. 2 are given in App. A, with the operator definitions in App. B and the mapping of existing methods onto them in App. C. App. D documents the full experimental setup: the synthetic data-generating process, estimators, metrics, and Monte Carlo repetitions for the recovery study (App. D.1, Tabs. 3–4); the kernel definitions and selection rationale (App. D.2, Tab. 5); and, for both real-world applications, data sources, preprocessing, feature construction, splits, model and hyperparameter settings, cooperative-game estimation with all sample sizes and fixed seeds, and instance-selection criteria (App. D.3, App. D.4). Data availability is described in App. D.8: IHEPC and NESO are public, while the licensed SPY bars cannot be redistributed, so we release the precomputed cooperative-game caches from which all SPY figures can be regenerated. Computing environment and runtimes are reported in App. D.9. An anonymized implementation with scripts reproducing all figures and tables is available at `https://anonymous.4open.science/r/Hilbert_functional_decomposition-FD8F/README.md`.

6.1 AI Use Statement

In this work, we used generative AI tools to design and provide feedback on experiments, generate synthetic data sets, to implement and edit software code, to clean and reformat datasets, to create and modify scientific figures, to identify related work, and to draft and polish sections of the manuscript. We have not used generative AI tools to help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, implement methods, assist with translation, support qualitative and thematic data analysis or interpret results. We have reviewed all AI-assisted work: code was checked and executed by the authors and verified against previously computed results, synthetic data and figures were checked and modified by the authors to match intended specifications, suggested related work was verified, and all AI-drafted text was revised by the authors. We take

responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Alen Alexanderian, Pierre A. Gremaud, and Ralph C. Smith. Variance-based sensitivity analysis for time-dependent processes. *Reliability Engineering & System Safety*, 196:106722, 2020. doi: 10.1016/j.ress.2019.106722.

Torben G. Andersen and Tim Bollerslev. Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. *International Economic Review*, 39(4):885–905, 1998. doi: 10.2307/2527343.

Philippe Bracke, Anupam Datta, Cougar Jung, and Shayak Sen. Machine learning explainability in finance: An application to default risk analysis. *Bank of England Working Paper*, (816), 2019.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324.

Katherine Campbell, Michael D. McKay, and Brian J. Williams. Sensitivity analysis when model outputs are functions. *Reliability Engineering & System Safety*, 91(10–11):1468–1472, 2006. doi: 10.1016/j.ress.2005.11.049.

Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. Doctor AI: Predicting clinical events via recurrent neural networks. In *Machine Learning for Healthcare Conference*, pp. 301–318. PMLR, 2016.

Ian Covert, Scott M. Lundberg, and Su-In Lee. Understanding global feature contributions with additive importance measures. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Ian Covert, Scott Lundberg, and Su-In Lee. Explaining by removing: A unified framework for model explanation. *Journal of Machine Learning Research*, 22(209):1–90, 2021.

Huiqi Deng, Na Zou, Mengnan Du, Weifu Chen, Guocan Feng, Ziwei Yang, Zheyang Li, and Quanshi Zhang. Unifying fourteen post-hoc attribution methods with Taylor interactions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(7):4625–4640, 2024. doi: 10.1109/TPAMI.2024.3358410.

Piotr Donizy, Mateusz Krzyzinski, Anna Markiewicz, Pawel Karpinski, Krzysztof Kotowski, Artur Kowalik, Jolanta Orlowska-Heitzman, Bozena Romanowska-Dixon, Przemyslaw Biecek, and Mai P. Hoang. Machine learning models demonstrate that clinicopathologic variables are comparable to gene expression prognostic signature in predicting survival in uveal melanoma. *European Journal of Cancer*, 174:251–260, 2022. doi: 10.1016/j.ejca.2022.07.031.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2019. URL `https://archive.ics.uci.edu`.

Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177):1–81, 2019.

Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232, 2001. doi: 10.1214/aos/1013203451.

Fabian Fumagalli, Maximilian Muschalik, Patrick Kolpaczki, Eyke Hüllermeier, and Barbara Hammer. SHAP-IQ: Unified approximation of any-order Shapley interactions. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Fabian Fumagalli, Maximilian Muschalik, Eyke Hüllermeier, Barbara Hammer, and Julia Herbinger. Unifying feature-based explanations with functional ANOVA and cooperative game theory. In *International Conference on Artificial Intelligence and Statistics*, pp. 5140–5148. PMLR, 2025.

Fabrice Gamboa, Alexandre Janon, Thierry Klein, and Agnès Lagnoux. Sensitivity analysis for multidimensional and functional outputs. *Electronic Journal of Statistics*, 8(1):575–603, 2014. doi: 10.1214/14-EJS895.

Alex Goldstein, Adam Kapelner, Justin Bleich, and Emil Pitkin. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, 24(1):44–65, 2015. doi: 10.1080/10618600.2014.907095.

Michel Grabisch. k-order additive discrete fuzzy measures and their representation. *Fuzzy Sets and Systems*, 92(2):167–189, 1997. doi: 10.1016/S0165-0114(97)00168-1.

Michel Grabisch and Marc Roubens. An axiomatic approach to the concept of interaction among players in cooperative games. *International Journal of Game Theory*, 28(4):547–565, 1999. doi: 10.1007/s001820050125.

Juan-José Giraldo Gutierrez, Evelyn Lau, Subhashini Dharmapalan, Melody Parker, Yurui Chen, Mauricio A. Álvarez, and Dennis Wang. Multi-output prediction of dose–response curves enables drug repositioning and biomarker discovery. *npj Precision Oncology*, 8(1):209, 2024. doi: 10.1038/s41698-024-00691-x.

Julia Herbinger, Gabriel Laberge, Maximilian Muschalik, Yann Pequignot, Marvin N. Wright, and Fabian Fumagalli. GRANITE: A generalized regional framework for identifying agreement in feature-based explanations. *arXiv preprint arXiv:2601.22771*, 2026.

Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3):293–325, 1948. doi: 10.1214/aoms/1177730196.

Giles Hooker. Discovering additive structure in black box functions. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 575–580, 2004. doi: 10.1145/1014052.1014122.

Neil Jethani, Mukund Sudarshan, Ian Covert, Su-In Lee, and Rajesh Ranganath. FastSHAP: Real-time Shapley value estimation. In *International Conference on Learning Representations*, 2022.

Hachem Kadri, Emmanuel Duflos, Philippe Preux, Stéphane Canu, Alain Rakotomamonjy, and Julien Audiffren. Operator-valued kernels for learning from functional response data. *Journal of Machine Learning Research*, 17(20):1–54, 2016. URL https://jmlr.org/papers/volume17/11-315/11-315.pdf.

Tim Kreuzer, Jelena Zdravkovic, and Panagiotis Papapetrou. PAX-TS: Model-agnostic multi-granular explanations for time series forecasting via localized perturbations. *arXiv preprint arXiv:2508.18982*, 2025.

Mateusz Krzyziński, Mikołaj Spytek, Hubert Baniecki, and Przemysław Biecek. SurvSHAP(t): Time-dependent explanations of machine learning survival models. *Knowledge-Based Systems*, 262:110234, 2023. doi: 10.1016/j.knosys.2022.110234.

Matieyendou Lamboni, Hervé Monod, and David Makowski. Multivariate sensitivity analysis to measure global contribution of input factors in dynamic models. *Reliability Engineering & System Safety*, 96(4):450–459, 2011. doi: 10.1016/j.ress.2010.12.002.

Sophie Hanna Langbein, Hubert Baniecki, Fabian Fumagalli, Niklas Koenen, Marvin N. Wright, and Julia Herbinger. Functional decomposition and Shapley interactions for interpreting survival models. *arXiv preprint arXiv:2602.16505*, 2026.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Daniel Lundstrom and Meisam Razaviyayn. A unifying framework to the analysis of interaction methods using synergy functions. In *International Conference on Machine Learning*, pp. 23005–23032. PMLR, 2023.

National Energy System Operator. Historic demand data. https://www.neso.energy/data-portal/historic-demand-data, 2024. Accessed: 2024.

Alexander Neubauer, Stefan Brandt, and Martin Kriegel. Explainable multi-step heating load forecasting: Using SHAP values and temporal attention mechanisms for enhanced interpretability. *Energy and AI*, 20:100480, 2025. doi: 10.1016/j.egyai.2025.100480.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Polygon.io. SPY intraday bar data. `https://polygon.io`, 2024. Five-minute OHLCV bars, accessed January 2022 – April 2024.

Vikas C. Raykar, Arindam Jati, Sumanta Mukherjee, Nupur Aggarwal, Kanthi Sarpatwar, Giridhar Ganapavarapu, and Roman Vaculin. TsSHAP: Robust model agnostic feature-based explainability for time series forecasting. *arXiv preprint arXiv:2303.12316*, 2023.

Gian-Carlo Rota. On the foundations of combinatorial theory: I. Theory of Möbius functions. In *Classic Papers in Combinatorics*, pp. 332–360. Springer, 1964.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019. doi: 10.1038/s42256-019-0048-x.

David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3): 1181–1191, 2020. doi: 10.1016/j.ijforecast.2019.07.001.

Andrea Saltelli, Marco Ratto, Terry Andres, Francesca Campolongo, Jessica Cariboni, Debora Gatelli, Michaela Saisana, and Stefano Tarantola. *Global Sensitivity Analysis: The Primer*. John Wiley & Sons, 2008.

Lloyd S. Shapley. A value for n-person games. *Contributions to the Theory of Games*, 2(28):307–317, 1953.

Yan Shi, Zhenzhou Lu, Zhao Li, and Mengmeng Wu. Cross-covariance based global dynamic sensitivity analysis. *Mechanical Systems and Signal Processing*, 100:846–862, 2018. doi: 10.1016/j.ymssp.2017.08.013.

Ilya M. Sobol'. Sensitivity estimates for nonlinear mathematical models. *Mathematical Modeling and Computational Experiment*, 1:407–414, 1993.

Ilya M. Sobol'. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Mathematics and Computers in Simulation*, 55(1–3):271–280, 2001. doi: 10.1016/S0378-4754(00)00270-6.

Carolin Strobl, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. Conditional variable importance for random forests. *BMC Bioinformatics*, 9:1–11, 2008. doi: 10.1186/1471-2105-9-307.

Erik Štrumbelj and Igor Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, 41(3):647–665, 2014. doi: 10.1007/s10115-013-0679-x.

Erico Tjoa and Cuntai Guan. A survey on explainable artificial intelligence (XAI): Towards medical XAI. *IEEE Transactions on Neural Networks and Learning Systems*, 32(11):4793–4813, 2021. doi: 10.1109/TNNLS.2020.3027314.

Corne van Zyl, Xianming Ye, and Raj Naidoo. Harnessing explainable artificial intelligence for feature selection in time series energy forecasting: A comparative analysis of Grad-CAM and SHAP. *Applied Energy*, 353:122079, 2024. doi: 10.1016/j.apenergy.2023.122079.

Yuyi Zhang, Qiushi Sun, Dongfang Qi, Jing Liu, Ruimin Ma, and Ovanes Petrosian. ShapTime: A general XAI approach for explainable time series forecasting. In *Proceedings of SAI Intelligent Systems Conference*, pp. 659–673. Springer, 2023. doi: 10.1007/978-3-031-47721-8_45.

APPENDIX

A PROOFS

Here, we provide the proofs for the theoretical statements in Sec.4.

A.1 PROOF OF THM. 1

Proof. By definition of the time-specific output behavior operator for prediction-based input behavior $\nu_{\text{inp}}(S)(t) = F_S^{\mathcal{H}}(\mathbf{x})(t)$,

$$(\mathcal{B}_{\text{out}}^{\text{spec}}\nu_{\text{inp}})(S,t) = \int_{\mathcal{T}} K(t,s)\,\nu_{\text{inp}}(S)(s)\,ds.$$

Assuming the integrals are well-defined, Fubini's theorem yields

$$\begin{aligned}(\mathcal{B}_{\text{out}}^{\text{aggr}}\nu_{\text{inp}})(S) &= \int_{\mathcal{T}} (\mathcal{B}_{\text{out}}^{\text{spec}}\nu_{\text{inp}})(S,t)\,dt \\ &= \int_{\mathcal{T}}\int_{\mathcal{T}} K(t,s)\,\nu_{\text{inp}}(S)(s)\,ds\,dt \\ &= \int_{\mathcal{T}} \left(\int_{\mathcal{T}} K(t,s)\,dt\right) \nu_{\text{inp}}(S)(s)\,ds.\end{aligned}$$

Defining

$$w_K(s) := \int_{\mathcal{T}} K(t,s)\,dt,$$

we obtain the stated representation.

For the ordering, consider individual pure feature effects and let $i, j \in P := \{1,\dots,p\}$. If $\mathcal{B}_{\text{inp}}$ is **linear** in $F^{\mathcal{H}}$ and

$$\nu_{\text{inp}}(\{i\})(s) \geq \nu_{\text{inp}}(\{j\})(s) \qquad \text{for all } s \in \mathcal{T},$$

then, since $K(t,s) \geq 0$, it follows that

$$w_K(s) = \int_{\mathcal{T}} K(t,s)\,dt \geq 0 \qquad \text{for all } s \in \mathcal{T}.$$

Hence,

$$w_K(s)\big(\nu_{\text{inp}}(\{i\})(s) - \nu_{\text{inp}}(\{j\})(s)\big) \geq 0 \qquad \text{for all } s \in \mathcal{T},$$

and therefore

$$\int_{\mathcal{T}} w_K(s)\big(\nu_{\text{inp}}(\{i\})(s) - \nu_{\text{inp}}(\{j\})(s)\big)\,ds \geq 0.$$

Using the representation above, this yields

$$(\mathcal{B}_{\text{out}}^{\text{aggr}}\nu_{\text{inp}})(\{i\}) \geq (\mathcal{B}_{\text{out}}^{\text{aggr}}\nu_{\text{inp}})(\{j\}),$$

which proves the result for individual pure feature effects.

Extension to partial and full effects. For partial effects (Shapley values), the attribution of a feature i is given by

$$\phi_i^{\text{Shap}} = \sum_{S \subseteq P\setminus\{i\}} \alpha_S\big[\nu(S\cup\{i\}) - \nu(S)\big],$$

where $\alpha_S > 0$ are the Shapley weights.

To compare two features i and j, consider the difference

$$\phi_i^{\text{Shap}} - \phi_j^{\text{Shap}}.$$

Substituting the definition and collecting terms yields

$$\phi_i^{\text{Shap}} - \phi_j^{\text{Shap}} = \sum_{S \subseteq P\setminus\{i\}} \alpha_S\big[\nu(S\cup\{i\}) - \nu(S)\big] - \sum_{S \subseteq P\setminus\{j\}} \alpha_S\big[\nu(S\cup\{j\}) - \nu(S)\big].$$

By reindexing the sums and grouping terms over coalitions $A \subseteq P \setminus \{i, j\}$, this expression can be written as a sum of terms of the form

$$\nu(A \cup \{i\}) - \nu(A \cup \{j\}),$$

each with a positive weight.

Hence, if for all $A \subseteq P \setminus \{i, j\}$ and all $s \in \mathcal{T}$,

$$\nu_{\text{inp}}(A \cup \{i\})(s) \geq \nu_{\text{inp}}(A \cup \{j\})(s),$$

then, by the same argument as in the main proof,

$$\nu(A \cup \{i\}) - \nu(A \cup \{j\}) = \int_{\mathcal{T}} w_K(s) \big[\nu_{\text{inp}}(A \cup \{i\})(s) - \nu_{\text{inp}}(A \cup \{j\})(s)\big] ds \geq 0,$$

and therefore $\phi_i^{\text{Shap}} \geq \phi_j^{\text{Shap}}$.

For full effects,

$$\phi_i^{\text{full}} = \nu(P) - \nu(P \setminus \{i\}),$$

so that

$$\phi_i^{\text{full}} - \phi_j^{\text{full}} = \nu(P \setminus \{j\}) - \nu(P \setminus \{i\}).$$

This corresponds to the same condition with $A = P \setminus \{i, j\}$, since

$$P \setminus \{j\} = A \cup \{i\}, \quad P \setminus \{i\} = A \cup \{j\}.$$

Hence, if

$$\nu_{\text{inp}}(A \cup \{i\})(s) \geq \nu_{\text{inp}}(A \cup \{j\})(s) \quad \forall s \in \mathcal{T},$$

then $\phi_i^{\text{full}} \geq \phi_j^{\text{full}}$.

□

A.2 Proof of Proposition 1

Proof. By construction,

$$\nu : 2^{\{1,\dots,p\}} \to \mathbb{R}, \qquad S \mapsto (\mathcal{B}_{\text{out}}(\mathcal{B}_{\text{inp}}(\mathcal{M}F^{\mathcal{H}})))(S),$$

is a well-defined set function on the power set $2^{\{1,\dots,p\}}$.

Let m_ν denote its Möbius transform, defined by

$$m_\nu(S) := \sum_{L \subseteq S} (-1)^{|S|-|L|} \nu(L),$$

which captures the contributions of individual subsets via inclusion–exclusion.

By Möbius inversion (Grabisch, 1997), ν admits the representation

$$\nu(T) = \sum_{S \subseteq T} m_\nu(S) \qquad \text{for all } T \subseteq \{1, \dots, p\}.$$

Applying this to $T = \{1, \dots, p\}$ and substituting the definition of m_ν yields

$$\nu(\{1, \dots, p\}) = \sum_{S \subseteq \{1,\dots,p\}} \sum_{L \subseteq S} (-1)^{|S|-|L|} \nu(L),$$

which is the claimed decomposition.

Since the Möbius transform is unique, this representation defines a unique additive decomposition of ν into contributions associated with feature subsets. As the construction of ν is independent of the specific choices of $\mathcal{M}$, $\mathcal{B}_{\text{inp}}$, and $\mathcal{B}_{\text{out}}$, the result holds for all corresponding time-specific, time-resolved, and time-aggregated explanations. □

A.3 Proof of Theorem 2

Proof. By definition of the sensitivity input behavior,

$$\nu_{\text{inp}}(S)(t,s) = \text{Cov}\big(F_S^{\mathcal{H}}(\mathbf{X})(t), F_S^{\mathcal{H}}(\mathbf{X})(s)\big).$$

For the constant kernel $K(t,s) = 1$, the output behavior in Eq. 6 reduces to integration over the second output argument:

$$(\mathcal{B}_{\text{out}}\nu_{\text{inp}}(S))(t) = \int_{\mathcal{T}} \nu_{\text{inp}}(S)(t,s)\, ds.$$

Thus,

$$\nu(S)(t) = \int_{\mathcal{T}} \text{Cov}\big(F_S^{\mathcal{H}}(\mathbf{X})(t), F_S^{\mathcal{H}}(\mathbf{X})(s)\big)\, ds,$$

and integrating once more over t yields

$$\nu(S) = \int_{\mathcal{T}} \nu(S)(t)\, dt = \int_{\mathcal{T}}\int_{\mathcal{T}} \nu_{\text{inp}}(S)(t,s)\, dt\, ds.$$

Under marginal masking,

$$F_S^{\mathcal{H}}(\mathbf{X}) = \mathbb{E}_{\mathbf{X}_{-S}\sim P_{\mathbf{X}_{-S}}}\left[F^{\mathcal{H}}(\mathbf{X}) \mid \mathbf{X}_S = \mathbf{x}_S\right],$$

so $\nu(S)(t)$ and $\nu(S)$ are exactly the covariance-based Sobol value functions for time-specific and time-aggregated trajectory outputs.

Under the feature independence assumption, the Hilbert-valued functional decomposition yields orthogonal pure effects $f_L^{\mathcal{H}}$, and therefore

$$\nu(S)(t) = \sum_{L\subseteq S}\int_{\mathcal{T}} \text{Cov}\big(f_{L,t}^{\mathcal{H}}(\mathbf{X}), f_{L,s}^{\mathcal{H}}(\mathbf{X})\big)\, ds.$$

Thus, the pure interaction operator, which aggregates contributions over $L \subseteq S$, recovers the closed Sobol contribution. Analogously, the full interaction operator aggregates all contributions involving at least one feature in S, i.e. all L with $L \cap S \neq \emptyset$, and therefore recovers the total Sobol contribution.

Normalizing these quantities by the corresponding total covariance quantity,

$$\nu(\{1,\ldots,p\})(t) = \int_{\mathcal{T}} \text{Cov}\big(F_t^{\mathcal{H}}(\mathbf{X}), F_s^{\mathcal{H}}(\mathbf{X})\big)\, ds,$$

gives the closed and total time-specific Sobol indices. Integrating the same covariance contributions over $t \in \mathcal{T}$ yields the corresponding time-aggregated closed and total Sobol indices. Hence, the proposed framework recovers these Sobol indices as a special case. □

B Operator Choices in the Hilbert-Valued Explanation Framework

We provide additional details on the operator choices underlying the Hilbert-valued explanation framework. Building on Fumagalli et al. (2025); Herbinger et al. (2026), we extend masking, behavior, and interaction operators to Hilbert-valued outputs by decomposing the behavior operator into input and output components, $\mathcal{B}_{\text{inp}}$ and $\mathcal{B}_{\text{out}}$.

Masking operators. Masking operators $\mathcal{M}$ define how features outside a subset S are removed by different masking strategies. Baseline masking fixes features to a reference value b_{-S}, marginal masking integrates them out under the marginal distribution, and conditional masking preserves all feature dependencies. Under independence, conditional and marginal masking coincides with the functional decomposition used in sensitivity analysis (functional ANOVA) (Sobol', 2001; Hooker, 2004).

Input behavior operators. The input behavior operator $\mathcal{B}_{\text{inp}}$ defines the quantity of interest prior to aggregation across the output domain. Prediction-based behavior captures local effects, sensitivity-based behavior captures variability via covariance, and risk-based behavior captures expected loss. Depending on the choice, $\nu_{\text{inp}}(S)$ is defined either pointwise over $\mathcal{T}$ or over pairs (t,s).

Table 1: Operator choices in the Hilbert-valued explanation framework.

Concept	Operator	Definition
Masking Operators ($\mathcal{M}^{(\cdot)}$)		
baseline	$\mathcal{M}^b$	$F_S^{\mathcal{H}}(\mathbf{x}) = F^{\mathcal{H}}(\mathbf{x}_S, b_{-S})$
marginal	$\mathcal{M}^m$	$F_S^{\mathcal{H}}(\mathbf{x}) = \mathbb{E}_{\mathbf{X}_{-S}}[F^{\mathcal{H}}(\mathbf{x}_S, \mathbf{X}_{-S})]$
conditional	$\mathcal{M}^c$	$F_S^{\mathcal{H}}(\mathbf{x}) = \mathbb{E}_{\mathbf{X}_{-S}\mid\mathbf{X}_S=\mathbf{x}_S}[F^{\mathcal{H}}(\mathbf{x}_S, \mathbf{X}_{-S})]$
Input Behavior Operators ($\mathcal{B}_{\text{inp}}^{(\cdot)}$)		
prediction	$\mathcal{B}_{\text{inp}}^{\text{pred}}$	$\nu_{\text{inp}}^{\text{pred}}(S)(t) = F_S^{\mathcal{H}}(\mathbf{x})(t)$
sensitivity	$\mathcal{B}_{\text{inp}}^{\text{sens}}$	$\nu_{\text{inp}}^{\text{sens}}(S)(t,s) = \text{Cov}(F_S^{\mathcal{H}}(\mathbf{X})(t), F_S^{\mathcal{H}}(\mathbf{X})(s))$
risk	$\mathcal{B}_{\text{inp}}^{\text{risk}}$	$\nu_{\text{inp}}^{\text{risk}}(S)(t) = \mathbb{E}[\ell(\mathbf{Y}(t), F_S^{\mathcal{H}}(\mathbf{X})(t))]$
Output Behavior Operators ($\mathcal{B}_{\text{out}}^{(\cdot)}$)		
time-specific	$\mathcal{B}_{\text{out}}^{\text{spec}}$	$\nu^{\text{spec}}(S,t) = \int K(t,s)\, g_S(s)\, ds$
time-resolved	$\mathcal{B}_{\text{out}}^{\text{res}}$	$\nu^{\text{res}}(S) = \{\nu^{\text{spec}}(S,t)\}_{t\in\mathcal{T}}$
time-aggregated	$\mathcal{B}_{\text{out}}^{\text{aggr}}$	$\nu^{\text{aggr}}(S) = \int\int K(t,s)\, g_S(s)\, ds\, dt$
Interaction Operators ($\mathcal{I}^{(\cdot)}$)		
pure	$\mathcal{I}^{\text{pure}}$	$\phi^{\text{pure}}(j) = \nu(j) - \nu(\emptyset)$
partial	$\mathcal{I}^{\text{partial}}$	$\phi^{\text{partial}}(j) = \sum_{S\subseteq -j} \frac{1}{p\binom{p-1}{\lvert S\rvert}}[\nu(S\cup j) - \nu(S)]$
full	$\mathcal{I}^{\text{full}}$	$\phi^{\text{full}}(j) = \nu(D) - \nu(-j)$

Output behavior operators. The output behavior operator $\mathcal{B}_{\text{out}}$ determines how Hilbert-valued quantities are aggregated across the output domain. Time-specific operators yield explanations at a fixed output location, time-resolved operators evaluate time-specific explanations for all $t \in \mathcal{T}$, and time-aggregated operators summarize contributions over the entire domain $\mathcal{T}$ into a scalar. All variants depend on a kernel K, which controls how dependencies across the output domain are incorporated.

Interaction operators. Interaction operators map set functions to feature-based explanations. Pure effects ignore interactions, partial effects (Shapley values) fairly distribute interaction effects across involved features, and full effects assign the entire interaction effect to each involved feature. Under marginal masking and sensitivity behavior, pure and full effects recover closed and total Sobol indices, respectively.

Method mapping. Many existing explanation methods arise as specific choices of masking, input behavior, and interaction operators (see Tab. 2). Prediction-based methods such as PDP and SHAP differ only in the interaction operator, while sensitivity-based methods correspond to Sobol indices with pure and full interactions yielding closed and total effects. Risk-based methods such as PFI and SAGE quantify performance degradation, differing again only in the interaction operator. The output behavior operator $\mathcal{B}_{\text{out}}$ determines based on a kernel choice K how these quantities are represented across the output domain (time-specific, time-resolved, or time-aggregated) and can be chosen independently, yielding valid explanations in all cases due to the additive decomposition established in Thm. 1. Detailed derivations of various feature-based explanation methods for the scalar case are given in Fumagalli et al. (2025).

C Mapping Existing Methods to the Hilbert-valued Explanation Framework

This section formalizes how existing feature-based explanation methods for time-dependent outputs can be expressed within the proposed operator framework. Throughout, we consider explanations of a Hilbert-valued prediction function $F^{\mathcal{H}}$ based on masking $\mathcal{M}$, input behavior $\mathcal{B}_{\text{inp}}$, output behavior $\mathcal{B}_{\text{out}}$, and interaction operator $\mathcal{I}$.

Table 2: Mapping of common explanation methods mentioned in this work to operator choices in the Hilbert-valued explanation framework. Output behavior $\mathcal{B}_{\text{out}}$ can be chosen independently (time-specific, time-resolved, or time-aggregated based on different kernel choices K), yielding valid explanations in all cases.

Method	$\mathcal{M}$	$\mathcal{B}_{\text{inp}}$	$\mathcal{I}$	Description
PDP	marginal	prediction	pure	average marginal effect
c-PDP	conditional	prediction	pure	average conditional marginal effect
SHAP (int.)	marginal	prediction	partial	(interventional) Shapley values
SHAP (obs.)	conditional	prediction	partial	(observational) Shapley values
Closed Sobol	marginal	sensitivity	pure	variance explained by features in S
Total Sobol	marginal	sensitivity	full	variance explained by features in S including all their interactions with $-S$
PFI	marginal	risk	full	model performance drop when removing feature j and all its interactions with $-j$ (marginal)
c-PFI	conditional	risk	full	model performance drop when removing feature j and all its interactions with $-j$ (conditional)
SAGE	marginal / conditional	risk	partial	global Shapley values based on risk games

Time-specific and time-aggregated output behavior. For prediction-based explanations, the input behavior is given by

$$\nu_{\text{inp}}(S)(t) = F_S^{\mathcal{H}}(\mathbf{X})(t).$$

Time-specific explanations correspond to

$$(\mathcal{B}_{\text{out}}^{\text{spec}} \nu_{\text{inp}})(S, t) = \nu_{\text{inp}}(S)(t),$$

while time-aggregated explanations are defined as

$$(\mathcal{B}_{\text{out}}^{\text{aggr}} \nu_{\text{inp}})(S) = \int_{\mathcal{T}} \nu_{\text{inp}}(S)(t)\, dt.$$

All existing methods considered here implicitly assume $K = I$, i.e., they do not model dependencies across the output domain and are the labeled pointwise baselines in Sec. 5.2.

If $\mathcal{B}_{\text{inp}}$ is linear in $F^{\mathcal{H}}$ and the aggregation operator defining $\mathcal{B}_{\text{out}}^{\text{aggr}}$ is linear (e.g., integration or averaging), then aggregation commutes with masking. Formally,

$$\int_{\mathcal{T}} F_S^{\mathcal{H}}(\mathbf{X})(t)\, dt = \mathbb{E}\left[\int_{\mathcal{T}} F^{\mathcal{H}}(\mathbf{X})(t)\, dt \;\middle|\; \mathbf{X}_S \right].$$

Hence, aggregating masked predictions is equivalent to computing explanations for an aggregated output (e.g., the mean prediction).

In practice, existing methods often implement aggregation by first defining a scalar output $\widetilde{F}$ (e.g., the mean forecast) and then applying standard explanation methods. Within our framework, this corresponds to treating aggregation as part of the output behavior operator.

This equivalence holds for linear aggregation functionals but not in general for nonlinear ones (e.g., maximum or quantiles), where aggregation and masking do not commute.

TsSHAP. TsSHAP corresponds to prediction-based explanations $\mathcal{B}_{\text{inp}}^{\text{pred}}$, partial interaction operator $\mathcal{I}^{\text{partial}}$, and time-specific output behavior

$$\mathcal{B}_{\text{out}} = \mathcal{B}_{\text{out}}^{\text{spec}}.$$

Explanations are computed independently for each $t \in \mathcal{T}$.

ShapTime. ShapTime explains aggregated forecast quantities (e.g., the mean prediction over the horizon. In practice, this is achieved by defining an aggregated scalar output prior to masking and applying standard Shapley methods. Within our framework, this corresponds to prediction-based input behavior $\mathcal{B}_{\text{inp}}^{\text{pred}}$ and time-aggregated output behavior combined with partial interaction operator $\mathcal{I}^{\text{partial}}$. For linear aggregation (e.g., the mean), both formulations are equivalent.

PAX-TS. PAX-TS computes feature importance via localized perturbations, measuring changes in selected output quantities (e.g., individual time points or aggregated statistics) when features are modified. Within our framework, this corresponds to prediction-based input behavior $\mathcal{B}_{\text{inp}}^{\text{pred}}$ combined with either time-specific or time-aggregated output behavior,

$$\mathcal{B}_{\text{out}} \in \{\mathcal{B}_{\text{out}}^{\text{spec}}, \mathcal{B}_{\text{out}}^{\text{aggr}}\},$$

and the full effects interaction operator $\mathcal{I}^{\text{full}}$. For linear aggregation functionals, this is equivalent to aggregating time-specific explanations.

SurvSHAP(t). SurvSHAP(t) provides time-specific and time-resolved explanations for survival models by computing Shapley values independently for each $t \in \mathcal{T}$. Within our framework, this corresponds to prediction-based input behavior $\mathcal{B}_{\text{inp}}^{\text{pred}}$, partial interaction operator $\mathcal{I}^{\text{partial}}$, and time-specific output behavior

$$\mathcal{B}_{\text{out}} = \mathcal{B}_{\text{out}}^{\text{spec}},$$

evaluated for all t, yielding a time-resolved explanation over the output domain.

SurvSHAP-IQ. SurvSHAP-IQ extends SurvSHAP(t) to interaction effects. In our framework, this corresponds to using higher-order interaction operators $\mathcal{I}$ (beyond partial effects), applied to the same time-specific output behavior (see Fumagalli et al. (2025) for definitions of higher-order interaction operators).

Remarks on masking. Some of the above methods, such as PAX-TS, consider time-dependent inputs and employ localized or time-dependent perturbations. While our framework currently focuses on marginal or conditional feature masking, these approaches can be interpreted as alternative masking strategies acting on structured or time-dependent inputs. Extending the framework to explicitly account for time-dependent masking operators is a natural direction for future work.

Summary. All considered methods correspond to prediction-based input behavior combined with either time-specific or time-aggregated output behavior under $K = I$, i.e., they treat time points independently and do not model dependencies across the output domain. In contrast, the proposed framework generalizes these approaches by allowing non-trivial kernels K for correlation-aware explanations and by incorporating alternative input behaviors (e.g., risk or sensitivity) for global importance measures.

D Experiments

This appendix contains full experimental details for all synthetic and real-data experiments in the paper. All analysis code is available at GitHub.

D.1 Ground-truth Recovery

Experimental setup

We validate the $\mathcal{H}$-FD framework on an extended version of the synthetic ICU biomarker model (Fig. 1), augmented with a pairwise interaction term to enable interaction recovery assessment:

$$F(\mathbf{x})(t) = X_1 e^{-0.2t} + X_2 e^{-(t-10)^2/2} + X_3 e^{-(t-18)^2/2} + \alpha(x_1 - \mu)(x_2 - \mu)\, e^{-(t-5)^2/2}, \quad (7)$$

with $X_i \sim \text{Uniform}[0, 1]$, $\mu = 0.5$, $\alpha = 1.0$, evaluated at $\boldsymbol{x}^* = (0.8,\ 0.9,\ 0.7)$. The interaction term accounts for approximately 2.4% of total variance. Gaussian observation noise with signal-to-noise ratio of 5 is added during training.

We estimate pure prediction effects via the Möbius transform applied to four black-box model classes fitted to n noisy observations: Ridge regression (deliberately misspecified, without interaction

term), random forest, NGBoost, and a multi-layer perceptron (MLP). An oracle estimator applies the Möbius transform directly to the true model, isolating estimation variance from model approximation error. Experiments are run across $n \in \{50, 100, 250, 500, 1000, 2000, 5000, 10000\}$ training samples, averaged over 30 Monte Carlo runs.

Evaluation metrics

Recovery is evaluated using two complementary metrics. Let $\hat{f}_S(t)$ denote the estimated pure effect and $f_S(t)$ the analytical ground truth.

Normalized L^2 error.

$$\varepsilon_{L^2}(S) = \frac{\|\hat{f}_S - f_S\|_{L^2}}{\|f_S\|_{L^2}} = \frac{\left(\int (\hat{f}_S(t) - f_S(t))^2 \, \mathrm{d}t\right)^{1/2}}{\left(\int f_S(t)^2 \, \mathrm{d}t\right)^{1/2}}, \tag{8}$$

measuring pointwise trajectory recovery relative to the signal magnitude. An error of 1.0 corresponds to a null (zero) predictor; errors exceeding 1.0 indicate the estimated effect is further from the truth than predicting zero, typically occurring when the wrong sign is learned at small n.

Relative aggregated error.

$$\varepsilon_{\mathrm{agg}}(S) = \frac{|\hat{\Phi}_S - \Phi_S|}{\Phi_S}, \qquad \Phi_S = \int |f_S(t)| \, \mathrm{d}t, \tag{9}$$

measuring recovery of the time-aggregated importance Φ_S, the quantity directly used for feature ranking in practice.

Effect recovery. Fig. 5 shows normalized L^2 error and relative aggregated error as a function of n for all four feature subsets. The oracle converges rapidly for all effects, confirming that the Möbius estimation procedure is consistent given the true model. All non-linear models recover the three main effects with L^2 errors below 5% at $n = 2{,}000$ and below 2% at $n = 10{,}000$; relative aggregated errors follow a similar pattern. The interaction term $f_{\{X_1,X_2\}}$ is harder to recover at all sample sizes due to its small signal contribution, but non-linear models consistently converge toward zero error with increasing data. Ridge regression recovers the main effects at a comparable rate but plateaus at $\varepsilon_{L^2} = 1.0$ and $\varepsilon_{\mathrm{agg}} = 1.0$ for the interaction at all n, correctly reflecting its structural inability to represent this term. Full numerical results are provided in Tabs. 3 and 4.

Time-resolved effect curves and time-aggregated effects. The top row of Fig. 6 shows representative time-resolved pure effect trajectories at $n = 1{,}000$. All non-linear models closely track the analytical ground truth for the three main effects. For the interaction $f_{\{X_1,X_2\}}$, the peak at $t \approx 5$h is recovered in shape but with larger pointwise variance across models, consistent with the higher L^2 errors observed at this sample size. The bottom row of Fig. 6 shows time-aggregated effects $\Phi_S = \int |f_S(t)| \, \mathrm{d}t$ for a representative run. All models reproduce the correct ordering $\Phi_{\{X_1\}} > \Phi_{\{X_2\}} > \Phi_{\{X_3\}} > \Phi_{\{X_1,X_2\}}$ and closely match the analytical values, with the largest relative deviations for the interaction term.

Sobol index recovery. Fig. 7 validates Thm. 2: under the constant kernel $K(t, s) = 1$ and sensitivity input behavior, $\mathcal{H}$-FD recovers the classical time-resolved and time-aggregated Sobol indices. The oracle estimator matches the analytical ground truth to within numerical precision across the full time domain. Non-linear models recover the indices accurately in the regions where each feature dominates, the exponential decay window for X_1, the peak at $t \approx 10$h for X_2, and the peak at $t \approx 18$h for X_3, but exhibit noisy deviations in near-zero regions where no feature dominates (row 1). These deviations arise from normalization: the time-resolved index divides by a near-zero denominator when all pure variance effects are small, amplifying estimation noise. This is an artifact of the ratio form of the Sobol index rather than a failure of effect recovery. Time-aggregated indices (row 2) are correctly recovered by all models, confirming that the integrated quantities are robust to pointwise normalization noise.

Aggregated ranking preservation. Thm. 1 establishes that the time-aggregated effect under $\mathcal{B}_{\mathrm{out}}$ admits the weighted-integral representation

$$\int_{\mathcal{T}} (\mathcal{B}_{\mathrm{out}} \, \nu_{\mathrm{inp}})(S)(t) \, \mathrm{d}t = \int_{\mathcal{T}} w_K(s) \, \nu_{\mathrm{inp}}(S)(s) \, \mathrm{d}s, \qquad w_K(s) := \int_{\mathcal{T}} K(t, s) \, \mathrm{d}t.$$

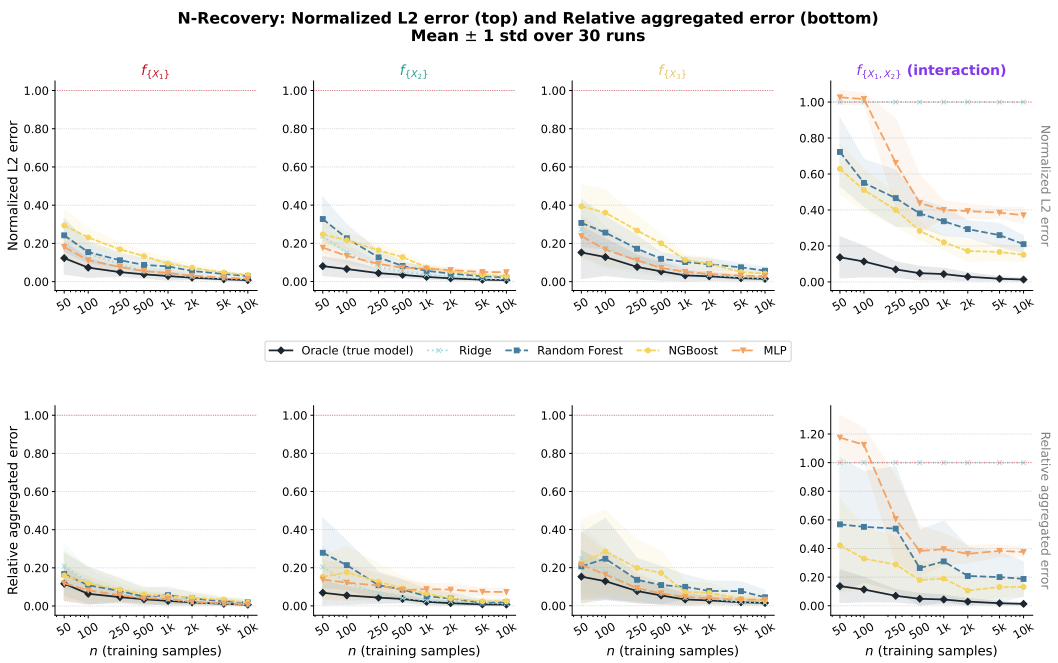


Figure 5: **N-Recovery**: Normalized L^2 error (*row 1*) and relative aggregated error (*row 2*) for the three main effects and the pairwise interaction $f_{\{X_1,X_2\}}$, as a function of training sample size n (mean $\pm$ 1 standard deviation (std) over 30 runs). The oracle lower bound reflects finite background sample noise. Main effects converge reliably across all models; the interaction term is substantially harder to recover, with the MLP showing the slowest convergence.

Table 3: Normalized L^2 error between estimated and oracle pure effects (mean $\pm$ 1 std over 30 independent runs). Each effect trajectory is normalized by the L^2 norm of the corresponding oracle effect, so a value of 1.0 corresponds to a trivially zero prediction. Ridge regression fails to recover the interaction effect $f_{\{X_1,X_2\}}$ at all sample sizes, reflecting model misspecification under the additive linear assumption.

Model / Effect	$n = 50$	$n = 100$	$n = 250$	$n = 500$	$n = 1$k	$n = 2$k	$n = 5$k	$n = 10$k
Oracle (true model)								
$f_{\{X_1\}}$	0.122 ± 0.068	0.062 ± 0.030	0.051 ± 0.031	0.043 ± 0.020	0.028 ± 0.019	0.025 ± 0.007	0.012 ± 0.006	0.008 ± 0.005
$f_{\{X_2\}}$	0.093 ± 0.045	0.074 ± 0.040	0.046 ± 0.022	0.034 ± 0.013	0.022 ± 0.010	0.016 ± 0.008	0.009 ± 0.005	0.007 ± 0.004
$f_{\{X_3\}}$	0.116 ± 0.112	0.124 ± 0.066	0.059 ± 0.048	0.049 ± 0.038	0.034 ± 0.036	0.024 ± 0.025	0.022 ± 0.020	0.014 ± 0.011
$f_{\{X_1,X_2\}}$	0.138 ± 0.101	0.135 ± 0.071	0.075 ± 0.043	0.068 ± 0.041	0.045 ± 0.029	0.031 ± 0.024	0.013 ± 0.011	0.013 ± 0.008
Ridge								
$f_{\{X_1\}}$	0.220 ± 0.092	0.131 ± 0.060	0.088 ± 0.035	0.059 ± 0.023	0.039 ± 0.021	0.030 ± 0.012	0.018 ± 0.006	0.013 ± 0.004
$f_{\{X_2\}}$	0.278 ± 0.067	0.169 ± 0.054	0.083 ± 0.020	0.056 ± 0.022	0.040 ± 0.014	0.026 ± 0.009	0.016 ± 0.004	0.013 ± 0.005
$f_{\{X_3\}}$	0.281 ± 0.126	0.202 ± 0.085	0.095 ± 0.047	0.067 ± 0.039	0.053 ± 0.036	0.034 ± 0.023	0.028 ± 0.020	0.017 ± 0.011
$f_{\{X_1,X_2\}}$	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
Random Forest								
$f_{\{X_1\}}$	0.248 ± 0.067	0.137 ± 0.027	0.105 ± 0.026	0.089 ± 0.038	0.070 ± 0.020	0.051 ± 0.018	0.042 ± 0.013	0.032 ± 0.009
$f_{\{X_2\}}$	0.387 ± 0.086	0.266 ± 0.071	0.135 ± 0.039	0.099 ± 0.029	0.063 ± 0.014	0.041 ± 0.010	0.027 ± 0.003	0.022 ± 0.003
$f_{\{X_3\}}$	0.317 ± 0.121	0.237 ± 0.077	0.142 ± 0.027	0.117 ± 0.023	0.105 ± 0.033	0.081 ± 0.027	0.083 ± 0.024	0.062 ± 0.018
$f_{\{X_1,X_2\}}$	0.816 ± 0.252	0.525 ± 0.095	0.405 ± 0.134	0.390 ± 0.058	0.358 ± 0.083	0.270 ± 0.047	0.234 ± 0.054	0.213 ± 0.057
NGBoost								
$f_{\{X_1\}}$	0.309 ± 0.116	0.237 ± 0.064	0.170 ± 0.027	0.151 ± 0.031	0.090 ± 0.019	0.067 ± 0.012	0.048 ± 0.011	0.036 ± 0.012
$f_{\{X_2\}}$	0.246 ± 0.036	0.214 ± 0.041	0.167 ± 0.032	0.133 ± 0.023	0.078 ± 0.024	0.054 ± 0.015	0.034 ± 0.010	0.031 ± 0.011
$f_{\{X_3\}}$	0.373 ± 0.118	0.368 ± 0.062	0.258 ± 0.052	0.199 ± 0.056	0.112 ± 0.022	0.093 ± 0.029	0.057 ± 0.022	0.042 ± 0.018
$f_{\{X_1,X_2\}}$	0.640 ± 0.126	0.506 ± 0.098	0.409 ± 0.096	0.274 ± 0.061	0.220 ± 0.033	0.184 ± 0.044	0.167 ± 0.047	0.158 ± 0.038
MLP								
$f_{\{X_1\}}$	0.186 ± 0.047	0.106 ± 0.039	0.080 ± 0.022	0.056 ± 0.020	0.042 ± 0.019	0.034 ± 0.013	0.023 ± 0.007	0.016 ± 0.005
$f_{\{X_2\}}$	0.211 ± 0.052	0.150 ± 0.048	0.094 ± 0.021	0.081 ± 0.014	0.075 ± 0.017	0.063 ± 0.011	0.050 ± 0.010	0.050 ± 0.005
$f_{\{X_3\}}$	0.203 ± 0.075	0.173 ± 0.064	0.095 ± 0.028	0.062 ± 0.022	0.052 ± 0.027	0.039 ± 0.017	0.032 ± 0.016	0.025 ± 0.010
$f_{\{X_1,X_2\}}$	1.036 ± 0.049	1.030 ± 0.031	0.527 ± 0.210	0.414 ± 0.177	0.388 ± 0.061	0.403 ± 0.036	0.378 ± 0.036	0.368 ± 0.044

For the prediction input behavior, $\mathcal{B}_{\text{inp}}$ is linear in $F^{\mathcal{H}}$, so if $\nu_{\text{inp}}(\{i\})(s) \geq \nu_{\text{inp}}(\{j\})(s)$ for all $s \in \mathcal{T}$, the integrated ordering is preserved for any non-negative kernel K: the kernel re-weights the time axis without reversing a pointwise-dominant feature. This sufficient condition holds in

Table 4: Relative aggregated error between estimated and oracle integrated pure effect values Φ_S (mean $\pm$ 1 std over 30 independent runs), normalized analogously to Tab. 3. For most estimator–subset combinations, aggregated errors tend to be lower than the corresponding L^2 errors due to partial cancellation of pointwise errors upon integration. For oracle estimates of $f_{\{X_3\}}$ and $f_{\{X_1,X_2\}}$, the two metrics coincide because the synthetic model is separable in those terms: the Möbius estimate inherits the true time-shape exactly, and finite-sample error reduces to a scalar coefficient that contributes identically to both metrics.

Model / Effect	$n = 50$	$n = 100$	$n = 250$	$n = 500$	$n = 1$k	$n = 2$k	$n = 5$k	$n = 10$k
Oracle (true model)								
$f_{\{X_1\}}$	0.117 ± 0.072	0.049 ± 0.033	0.047 ± 0.033	0.039 ± 0.021	0.025 ± 0.019	0.023 ± 0.007	0.012 ± 0.006	0.007 ± 0.005
$f_{\{X_2\}}$	0.083 ± 0.061	0.059 ± 0.038	0.049 ± 0.036	0.031 ± 0.017	0.018 ± 0.010	0.013 ± 0.010	0.008 ± 0.005	0.007 ± 0.004
$f_{\{X_3\}}$	0.116 ± 0.112	0.124 ± 0.066	0.059 ± 0.048	0.049 ± 0.038	0.034 ± 0.036	0.024 ± 0.025	0.022 ± 0.020	0.014 ± 0.011
$f_{\{X_1,X_2\}}$	0.138 ± 0.101	0.135 ± 0.071	0.075 ± 0.043	0.068 ± 0.041	0.045 ± 0.029	0.031 ± 0.024	0.013 ± 0.011	0.013 ± 0.008
Ridge								
$f_{\{X_1\}}$	0.193 ± 0.118	0.109 ± 0.071	0.075 ± 0.039	0.049 ± 0.036	0.029 ± 0.027	0.022 ± 0.016	0.012 ± 0.010	0.010 ± 0.007
$f_{\{X_2\}}$	0.217 ± 0.106	0.128 ± 0.073	0.075 ± 0.031	0.036 ± 0.030	0.032 ± 0.029	0.026 ± 0.012	0.017 ± 0.008	0.015 ± 0.013
$f_{\{X_3\}}$	0.326 ± 0.121	0.193 ± 0.129	0.066 ± 0.064	0.074 ± 0.043	0.059 ± 0.046	0.036 ± 0.025	0.030 ± 0.025	0.015 ± 0.016
$f_{\{X_1,X_2\}}$	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
Random Forest								
$f_{\{X_1\}}$	0.174 ± 0.100	0.083 ± 0.049	0.048 ± 0.047	0.039 ± 0.033	0.038 ± 0.025	0.038 ± 0.027	0.025 ± 0.017	0.025 ± 0.017
$f_{\{X_2\}}$	0.284 ± 0.212	0.240 ± 0.138	0.140 ± 0.067	0.113 ± 0.047	0.071 ± 0.035	0.043 ± 0.016	0.012 ± 0.012	0.025 ± 0.013
$f_{\{X_3\}}$	0.209 ± 0.205	0.176 ± 0.164	0.111 ± 0.063	0.108 ± 0.076	0.112 ± 0.057	0.077 ± 0.046	0.083 ± 0.050	0.069 ± 0.042
$f_{\{X_1,X_2\}}$	0.699 ± 0.592	0.551 ± 0.306	0.410 ± 0.380	0.286 ± 0.202	0.338 ± 0.329	0.191 ± 0.207	0.155 ± 0.103	0.207 ± 0.141
NGBoost								
$f_{\{X_1\}}$	0.180 ± 0.192	0.106 ± 0.061	0.072 ± 0.062	0.068 ± 0.040	0.054 ± 0.035	0.032 ± 0.026	0.032 ± 0.017	0.022 ± 0.018
$f_{\{X_2\}}$	0.111 ± 0.061	0.186 ± 0.152	0.140 ± 0.075	0.095 ± 0.061	0.080 ± 0.045	0.038 ± 0.027	0.025 ± 0.016	0.032 ± 0.022
$f_{\{X_3\}}$	0.246 ± 0.238	0.284 ± 0.179	0.224 ± 0.115	0.145 ± 0.115	0.079 ± 0.041	0.059 ± 0.055	0.044 ± 0.030	0.034 ± 0.025
$f_{\{X_1,X_2\}}$	0.394 ± 0.272	0.425 ± 0.224	0.240 ± 0.185	0.204 ± 0.105	0.217 ± 0.139	0.114 ± 0.086	0.125 ± 0.081	0.131 ± 0.068
MLP								
$f_{\{X_1\}}$	0.141 ± 0.096	0.073 ± 0.061	0.057 ± 0.038	0.039 ± 0.029	0.030 ± 0.027	0.025 ± 0.020	0.016 ± 0.010	0.010 ± 0.010
$f_{\{X_2\}}$	0.152 ± 0.123	0.131 ± 0.081	0.125 ± 0.042	0.091 ± 0.051	0.112 ± 0.026	0.096 ± 0.024	0.076 ± 0.021	0.072 ± 0.017
$f_{\{X_3\}}$	0.137 ± 0.125	0.161 ± 0.135	0.069 ± 0.054	0.048 ± 0.032	0.053 ± 0.044	0.035 ± 0.031	0.033 ± 0.018	0.026 ± 0.017
$f_{\{X_1,X_2\}}$	1.232 ± 0.192	1.162 ± 0.132	0.440 ± 0.257	0.325 ± 0.221	0.372 ± 0.135	0.361 ± 0.048	0.384 ± 0.046	0.374 ± 0.035

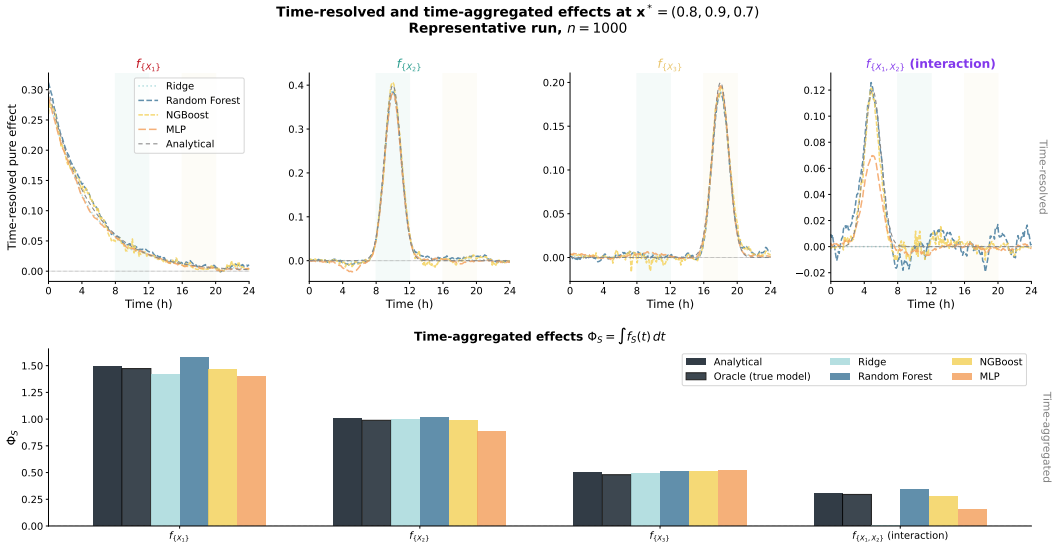


Figure 6: Time-resolved (*row 1*) and time-aggregated (*row 2*) $\mathcal{H}$-FD pure prediction effects at $\mathbf{x}^* = (0.8, 0.9, 0.7)$ for a representative run at $n = 1{,}000$. The time-resolved curves show estimated effects overlaid with the analytical ground truth (grey dashed). Main effects are closely recovered by all models; the interaction $f_{\{X_1,X_2\}}$ shows greater variability, and the MLP underestimates its time-aggregated value.

the three examples of main-text Fig. 2, where the time-aggregated rankings $\Phi_{X_1} > \Phi_{X_2} > \Phi_{X_3}$ (ICU), $\Phi_{X_2} > \Phi_{X_3} > \Phi_{X_1}$ (price pulse), and $\Phi_{X_1} \gg \Phi_{X_2}$ (periodic medication) are preserved across all kernels under the prediction input behavior. By contrast, Remark 1 (4.2) establishes that ranking invariance does not generally hold for nonlinear input behaviors. Although the same weighted-integral representation applies, different kernels may induce different feature rankings

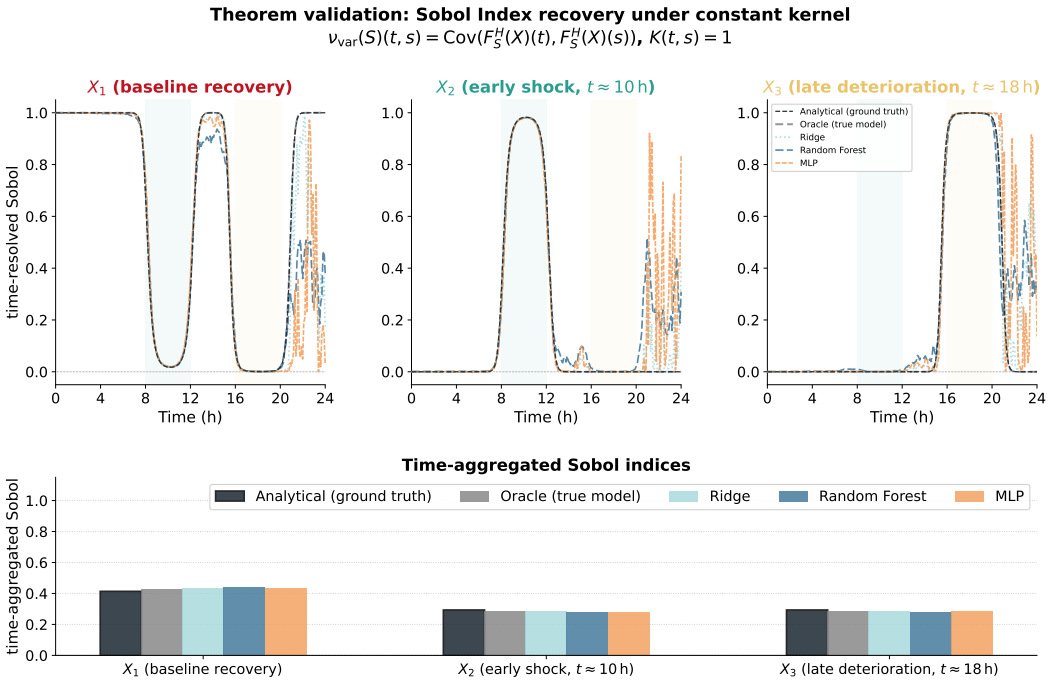


Figure 7: Time-resolved (*row 1*) and time-aggregated (*row 2*) Sobol indices recovered by $\mathcal{H}$-FD under the constant kernel and sensitivity input behavior, compared against the analytical ground truth.

under sensitivity or risk input behaviors. We illustrate this directly for the ICU model in Fig. 8: the prediction input behavior preserves the ordering $X_1 > X_2 > X_3$ across identity, OU, and correlation kernels, while the sensitivity and risk input behaviors produce a different ranking under different kernels $X_1 > X_3 > X_2$.

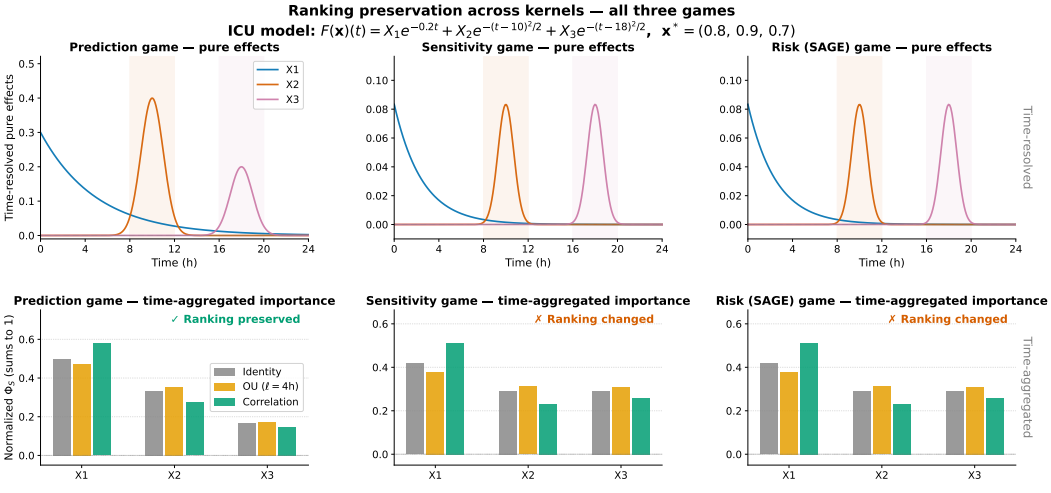


Figure 8: Ranking preservation across input behaviors for the ICU model, under identity, OU ($\ell = 4$h), and correlation kernels. (*row 1*) Time-resolved pure effects; (*row 2*) time-aggregated normalized importance. Under the prediction input behavior (*col. 1*), $X_1 > X_2 > X_3$ is preserved across all kernels, consistent with Thm. 1. Under the sensitivity and risk input behaviors (*col. 2,3*), rankings change across kernels: the correlation kernel reweights the squared per-feature effects so that X_3 overtakes X_2, while identity and OU preserve $X_2 > X_3$. Pointwise dominance of individual effects is not preserved under the quadratic input behaviors of variance and MSE, as stated in Remark 1 (4.2).

Table 5: **Kernel selection guide** for $\mathcal{H}$-FD. The appropriate kernel depends on the research question, output structure, and domain knowledge; there is no universally correct choice. For outputs mixing features with different temporal characteristics, kernels may be selected per-feature or combined.

Kernel	Temporal assumption	Use when	Domains
$\delta(t-s)$ (identity)	Locations independent	Instantaneous effects; recovers pointwise baselines	Any; default
1 (constant)	All interactions equally weighted	Trajectory-level importance; Sobol indices	Sensitivity analysis
$\mathrm{Corr}(F_t, F_s)$ (correlation)	Weighted by output co-movement	Features drive distinct trajectory shapes	Finance, electricity consumption
$\exp(-\lvert t-s\rvert/\ell)$ (OU)	Mean-reverting, local smoothing	Noisy outputs; sustained influence over a window	Biomedical, intraday volatility
$\exp(-(t-s)^2/2\sigma^2)$ (Gaussian)	Smooth symmetric neighbourhoods	Smooth outputs; no causal ordering required	Temperature, demand
$\rho^{\lvert t-s\rvert}$ (AR-type)	Discrete autoregressive dependence	Strong lag structure; discrete-time predictions	Economic time series
$\exp(-\lvert t-s\rvert/\ell)\,\mathbf{1}_{t\geq s}$ (causal)	Strict temporal ordering	Event-driven; causal interpretability required	Algorithmic trading
$\exp(-2\sin^2(\pi(t-s)/\tau)/\ell^2)$ (periodic)	Known periodicity; apply per-feature	Multiple full periods; genuinely recurring influence	Circadian, seasonal

D.2 Kernel Guidance

This appendix provides supplementary details for the kernel guidance discussion in the main text (Sec. 5.1, Fig. 2).

Kernel role across input behaviors. For input behaviors defined pointwise over the output domain—prediction $\nu_{\mathrm{inp}}^{\mathrm{pred}}(S)(t) = F_S^{\mathcal{H}}(\mathbf{x})(t)$ and risk $\nu_{\mathrm{inp}}^{\mathrm{risk}}(S)(t)$—the kernel introduces dependencies between output locations: it specifies how feature effects at location t incorporate information from other locations $s \in \mathcal{T}$. For variance-based explanations, the input behavior already captures dependencies between output locations through the covariance surface

$$\nu_{\mathrm{inp}}^{\mathrm{sens}}(S)(t,s) = \mathrm{Cov}\big(F_S^{\mathcal{H}}(\mathbf{X})(t),\, F_S^{\mathcal{H}}(\mathbf{X})(s)\big),$$

and the kernel determines how these covariance contributions are aggregated across time pairs. Classical functional Sobol indices correspond to the constant kernel $K(t,s) = 1$, which assigns equal weight to all covariance terms (Gamboa et al., 2014; Alexanderian et al., 2020).

Kernel selection. Tab. 5 summarizes the temporal assumption and intended use case for a selected set of useful kernels. The appropriate choice depends on the research question, output structure, and domain knowledge; there is no universally correct kernel. For outputs mixing features with different temporal characteristics, kernels may be selected per-feature or combined.

Feature heterogeneity. Using a single kernel imposes the same temporal structure across features, which can misattribute effects when feature behaviors differ. A pharmacokinetic response curve for some medication over 3 days demonstrates both the value and the risk of the periodic kernel (Fig. 9, col. 3): it correctly captures the daily recurrence of X_1 (8am dosing) but spuriously induces periodicity for X_2, a one-off event. Kernel choice should reflect feature-specific temporal behavior; in mixed settings, selecting kernels per-feature may be appropriate. This matters in practice, since recurring covariates are routinely mixed with event-driven ones.

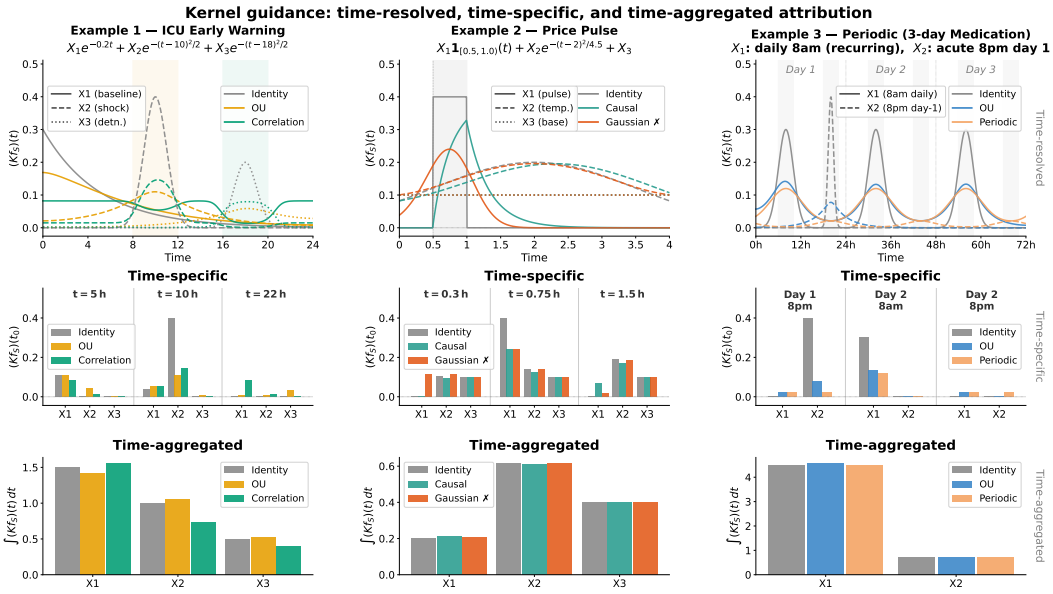


Figure 9: Kernel guidance for two synthetic examples with distinct temporal structures: an ICU early warning model (*col. 1*), a market price pulse (*col. 2*) and and a 3-day medication response (*col. 3*). Each example shows *pure local effects* at three explanation levels: time-resolved (*row 1*), time-specific at selected time points (*row 2*), and time-aggregated (*row 3*) for marginal masking $\mathcal{M}$, prediction-based input behavior $\mathcal{B}_{\text{inp}}$, and pure interaction operator $\mathcal{I}$ of a randomly chosen observation; the identity-kernel curves are the pointwise baseline under which existing methods operate (Sec. 4.3).

D.3 SPY INTRADAY VOLATILITY

This appendix provides full experimental details and supplementary figures for the SPY intraday volatility case study summarized in the main text (Sec.5.2, Fig. 3).

Data. We use one-minute OHLCV (open, high, low, close, volume) bar data for the SPDR S&P 500 ETF Trust (SPY) sourced from Polygon.io (`https://polygon.io`) (Polygon.io, 2024), covering 560 trading days from January 3, 2022 to April 30, 2024. The New York Stock Exchange regular session runs 09:30–16:00 ET; after removing the opening and closing auction bars we retain $T = 78$ five-minute intervals per day. The per-interval target is the *absolute log-return*

$$y_t \;=\; \bigl|\log(c_t/c_{t-1})\bigr|, \qquad t = 1, \ldots, 78,$$

where c_t denotes the close price of the t-th five-minute bar. Absolute log-returns are a standard proxy for intraday volatility at the five-minute frequency in the high-frequency literature (Andersen & Bollerslev, 1998). We subtract the cross-day mean at each interval (the *diurnal pattern*) so that the model is trained to explain day-to-day deviations from the typical U-shaped volatility curve (elevated at the open due to overnight news digestion, low through midday, and elevated again at the close due to position squaring), rather than the U-shape itself. The diurnal pattern is estimated on the training split only and then subtracted from both splits.

Pre-market features. Six scalar covariates are constructed from information available before the 09:30 open:

Feature	Description	Source
`vix_prev`	Prior business day's closing VIX level	CBOE via Yahoo Finance
`overnight_ret`	Log-return from prior close to 09:30 open	Polygon.io
`ann_indicator`	Binary: 1 if a scheduled macro announcement falls on this day[†]	Federal Reserve / BLS / BEA
`day_of_week`	Integer $\in \{0, \ldots, 4\}$ (Mon–Fri)	–
`trailing_rv`	Mean daily absolute log-return over the prior 5 trading days	Polygon.io
`month`	Integer $\in \{1, \ldots, 12\}$	–

[†] `ann_indicator` is set to 1 on FOMC meeting dates, CPI release dates, and Non-Farm Payroll (NFP) release dates; all dates are hardcoded from official Federal Reserve, BLS, and BEA schedules.

All continuous features (`vix_prev`, `overnight_ret`, `trailing_rv`) are standardized to zero mean and unit variance using statistics computed on the training split only.

Model. We use a single chronological train/test split to respect temporal ordering and prevent look-ahead leakage.

Split	Date range	# Days	%
Train	2022-01-03 – 2023-09-29	448	80%
Test	2023-10-02 – 2024-04-30	112	20%

We fit a *multivariate random forest*: a single `RandomForestRegressor` from `scikit-learn` (Pedregosa et al., 2011) whose targets are the $T = 78$ diurnal-adjusted volatility values stacked into a single matrix $\mathbf{Y} \in \mathbb{R}^{N \times T}$. Each tree predicts the full trajectory jointly, which naturally captures temporal correlations in the residuals without requiring an explicit sequence model.

Hyperparameters.

Hyperparameter	Value
`n_estimators`	300
`max_features`	`"sqrt"` (default)
`min_samples_leaf`	1 (default)
`min_samples_split`	2 (default)
`max_depth`	None (fully grown)
`bootstrap`	True (default)
`random_state`	42
`n_jobs`	-1 (all available cores)

Hyperparameters were fixed *a priori*; no grid search or cross-validation was performed. This choice was deliberate: our goal is to demonstrate the framework's explanatory properties rather than optimize predictive accuracy, and heavy tuning on 560 days of financial data risks over-fitting to the specific market regime covered. Random forests with 300 trees and default splitting rules are a well-understood baseline that generalizes reliably in low-sample tabular settings(Breiman, 2001). The model achieves a trajectory-level $R^2 = 0.120$ on the held-out test split ($1 - \mathrm{SS}_{\mathrm{res}}/\mathrm{SS}_{\mathrm{tot}}$ over all 112×78 predictions). This modest value is expected: the diurnal pattern has been subtracted prior to modelling, leaving only day-to-day deviations from the typical U-shape as the target.

Cooperative game set-up. The full 2^p-point Möbius transform is computed exactly over the Boolean lattice for all $S \subseteq [p]$. Pure main effects, partial (Shapley) effects, and full effects are then derived in closed form from the Möbius values. We use marginal (interventional) imputation throughout, sampling missing feature values from the training background set. All Monte Carlo estimators use a fixed random seed (seed $= 42$) for reproducibility.

The three input behaviors are estimated as follows.

Local prediction. For the High-VIX Announcement profile $\mathbf{x}^*$ and a coalition S,

$$\nu_{\mathrm{inp}}^{\mathrm{pred}}(S)(t) = \mathbb{E}_{\mathbf{X}_{-S}}\big[F(\mathbf{x}^*_S, \mathbf{X}_{-S})(t)\big],$$

estimated with $n_{\mathrm{sample}} = 200$ background draws per coalition. Pure effects yield per instance PDP equivalents, and partial effects are interventional SHAP values. The PDP-style plots in Fig. 13 aggregate per-instance local prediction effects over $N_{\mathrm{PDP}} = 120$ profiles drawn by stratified sampling from the training set (one instance per month plus random fill).

Global sensitivity.

$$\nu_{\mathrm{inp}}^{\mathrm{sens}}(S)(t,s) = \mathrm{Cov}\big(F_S^{\mathcal{H}}(\mathbf{X})(t),\, F_S^{\mathcal{H}}(\mathbf{X})(s)\big),$$

estimated by nested Monte Carlo: an outer sample of $n_{\mathrm{outer}} = 100$ draws of $\mathbf{X}_S$ from the training distribution, and for each outer draw an inner sample of $n_{\mathrm{inner}} = 100$ draws of $\mathbf{X}_{-S}$ for the conditional expectation. Pure effects yield closed Sobol indices and full effects yield total Sobol indices (Thm. 2); partial effects are the corresponding Shapley sensitivity indices.

Global risk. The risk effect is defined as the loss *reduction* relative to the marginal-only predictor $\bar{F}(t) = \mathbb{E}[F(\mathbf{X})](t)$:

$$\nu_{\mathrm{inp}}^{\mathrm{risk}}(S)(t) = \mathbb{E}\big[(Y(t) - \bar{F}(t))^2\big] - \mathbb{E}_{(\mathbf{X},Y)}\Big[\big(Y(t) - \mathbb{E}_{\mathbf{X}_{-S}}[F(\mathbf{X}_S, \mathbf{X}_{-S})(t)]\big)^2\Big],$$

estimated by the same nested Monte Carlo scheme with $n_{\mathrm{outer}} = n_{\mathrm{inner}} = 100$. This sign convention matches the SAGE / PFI literature, in which higher importance corresponds to larger reduction in expected loss; pure effects are also set to be non-negative by construction. With $p = 6$ features there are $2^p = 64$ coalitions per input behavior.

Kernels. Two kernel choices are compared throughout.

- **Identity kernel** $K = I$. Treats each time step independently; corresponds to the pointwise baselines (PD/TsSHAP, Sobol, PFI/SAGE, depending on $\mathcal{B}_{\mathrm{inp}}$).
- **Feature-specific kernel.** Rather than a single kernel for all features, we assign each feature a kernel that reflects its expected temporal structure. For all features except `ann_indicator`, we use an Ornstein–Uhlenbeck (OU) kernel,

  $$[K_{\mathrm{OU}}]_{st} = \exp\Big(-\tfrac{|t-s|}{\ell}\Big),$$

  with length-scale $\ell = 8$ (in units of five-minute bars, equivalent to 40 minutes), encoding smooth temporal decay. For `ann_indicator`, whose effect is expected to be localized to a specific event time with no pre-event leakage, we use a causal kernel,

  $$[K_{\mathrm{causal}}]_{st} = \exp\Big(-\tfrac{t-s}{\ell}\Big)\,\mathbf{1}[t \geq s],$$

  with the same length-scale $\ell = 8$. This enforces the interpretation that the announcement at time s can only affect attributions at $t \geq s$. The length-scale $\ell = 8$ was chosen to match the typical half-life of a volatility spike following a scheduled macro release (approximately

30–60 minutes in the high-frequency literature (Andersen & Bollerslev, 1998)), while remaining interpretable on the 78-bar trading day grid. Both kernels are row-normalized before application so that each time step receives a unit-mass weighted average of the effect trajectory.

Profile selection. The High-VIX Announcement profile studied in the main text (July 13, 2022) is selected as the median instance among all days satisfying $\texttt{ann_indicator} = 1$ and $\texttt{vix_prev} \geq q_{75}$, where q_{75} is the 75th percentile of vix_prev over the training set (22 days match these criteria). The profile has feature values $\texttt{vix_prev} = 27.29$, $\texttt{overnight_ret} = -0.0153$, $\texttt{ann_indicator} = 1$, $\texttt{day_of_week} = 2$ (Wednesday), $\texttt{trailing_rv} = 8.18 \times 10^{-4}$, and $\texttt{month} = 7$ (July).

Supplementary figures. Figs. 10–12 provide the full set of results for the SPY experiment. All panels use the following color coding: vix_prev (blue), ann_indicator (orange), overnight_ret (green), trailing_rv (brown), day_of_week (purple), month (red). Solid lines correspond to the OU kernel; dashed lines to the causal kernel (used for ann_indicator in the OU + causal kernel rows). The dashed vertical grey line marks 14:00 ET, the time of the Beige Book release on July 13, 2022.

Fig. 10 shows global sensitivity and risk effects, capturing how vix_prev contributes to output variance and forecast error across the full data distribution rather than at a single instance. vix_prev dominates both quantities across all effect types and both kernels, consistent with the well-established role of implied volatility as a forward-looking risk signal at the population level. Under the identity kernel its effect trajectory is noisy and roughly flat across the trading day, while the OU kernel reveals a smooth U-shape, reflecting that variance and forecast error are systematically elevated at the open and close. The gap between pure and full importances is largest for vix_prev, indicating that a substantial portion of its global contribution operates through interactions with other features rather than as a standalone main effect.

Fig. 11 shows local prediction effects for the High-VIX Announcement profile. Under the identity kernel, pure effects are noisy and difficult to interpret. The feature-specific kernel clarifies the structure: vix_prev produces a smooth positive baseline effect throughout the day, while the causal kernel localizes the effect of ann_indicator sharply after 14:00 with no pre-event leakage. The partial-to-pure ratio of $1.61\times$ for ann_indicator indicates that interaction with other features substantially amplifies its attributed effect.

Fig. 12 confirms that the $\texttt{vix_prev} \times \texttt{ann_indicator}$ pair is the strongest pairwise interactions in terms of time-aggregated effects. Under the identity kernel this manifests as a sharp spike at 14:00. The causal kernel smooths and causally localizes this spike, revealing the regime shift described in the main text: mild redundancy between the two features before the announcement, where both reflect elevated expected volatility, followed by strong synergy afterwards, where the release amplifies the impact of prior market fear. A second interaction, $\texttt{vix_prev} \times \texttt{overnight_ret}$, is comparable in time-aggregated magnitude but lacks a sharp event-locked profile; remaining pairs are noticeably smaller. Together this confirms that the vix_prev–ann_indicator interaction is the structurally dominant *event-driven* higher-order effect on this profile.

Fig. 13 shows that vix_prev exhibits a clear monotone relationship with predicted volatility deviations under the OU kernel: higher prior VIX is associated with larger positive effects across the full trading day, and this relationship is consistent across all five selected time points. Under the identity kernel the same relationship is present but noisier, with greater variability across time points. For overnight_ret the time-aggregated effect is near zero, suggesting that while the overnight return may matter for specific instances, its effect averages out across the training distribution.

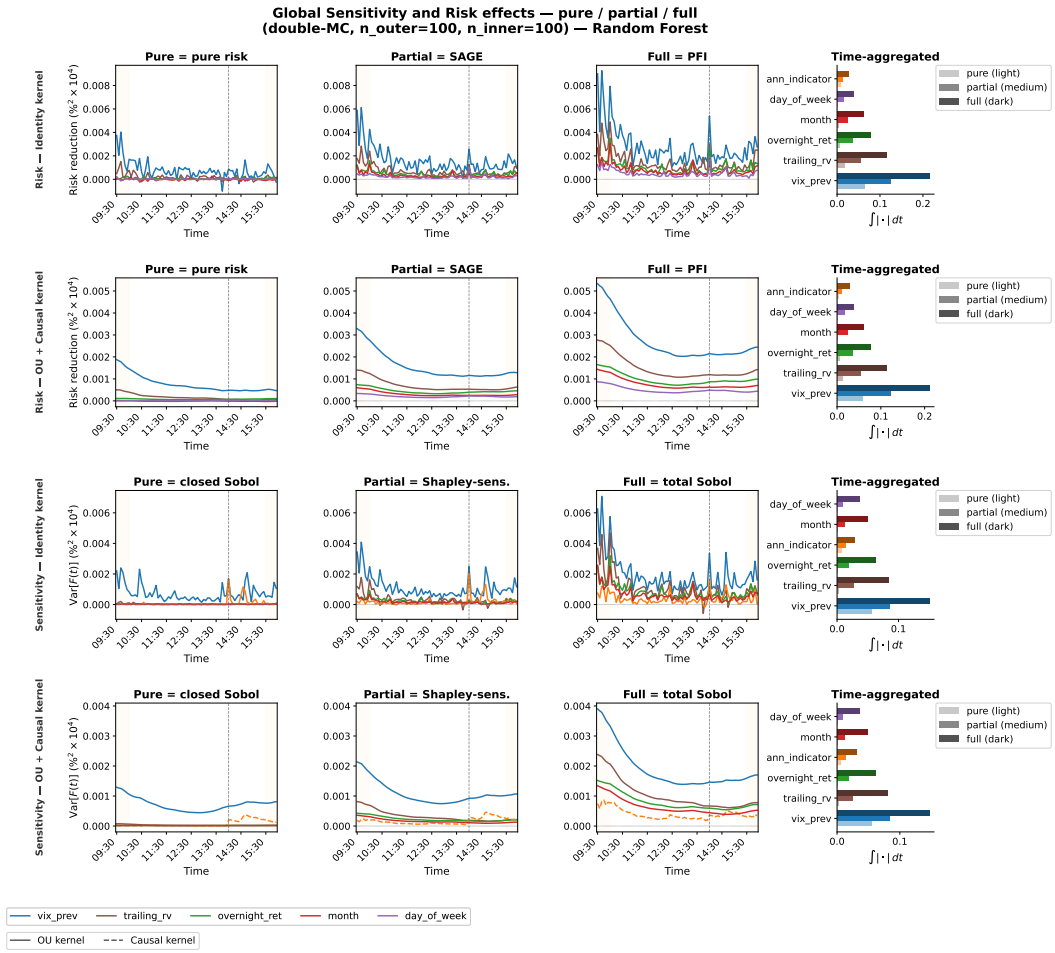


Figure 10: **Global sensitivity and risk effects — pure / partial / full.** Effects are computed by nested Monte Carlo ($n_{\text{outer}} = n_{\text{inner}} = 100$) directly from the global input behaviors $\nu_{\text{inp}}^{\text{sens}}$ and $\nu_{\text{inp}}^{\text{risk}}$. (*rows 1–2*) Risk behavior (loss reduction in $\%^2 \times 10^{-4}$) under the identity kernel (*row 1*) and feature-specific kernel (*row 2*); (*rows 3–4*) sensitivity input behavior under the identity (*row 3*) and feature-specific (*row 4*) kernels. (*cols. 1–3*) Pure = closed Sobol / pure risk (col. 1), partial = Shapley-sensitivity / SAGE (*col. 2*), full = total Sobol / PFI (*col. 3*). (*col. 4*) Time-aggregated bar chart for pure, partial, and full importances.

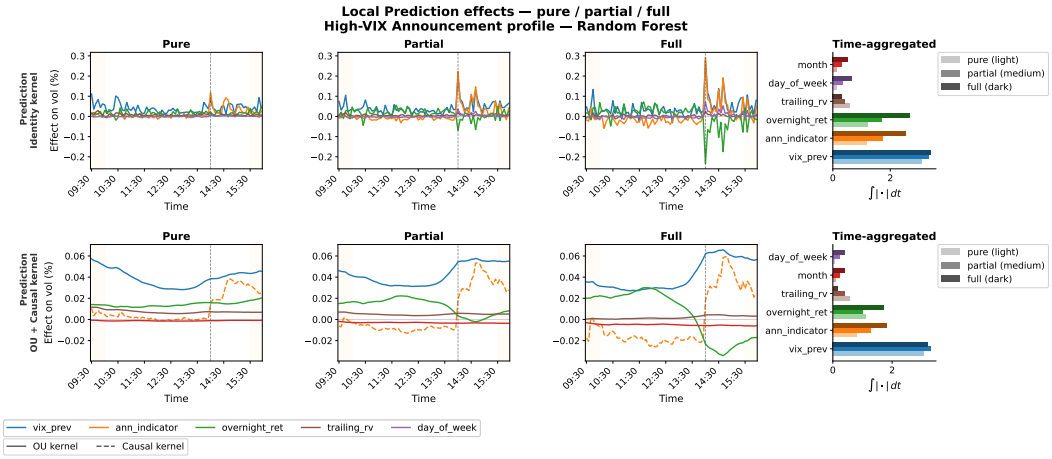


Figure 11: **Local prediction effects — pure / partial / full.** Instance: July 13, 2022 (`vix_prev` = 27.3, `ann_indicator` = 1), selected as the median among days with `ann_indicator` = 1 and `vix_prev` $\geq q_{75}$. (*Row 1*) Identity kernel (Pointwise PD/TsSHAP); (*Row 2*) feature-specific kernel (OU for all features except `ann_indicator`; causal for `ann_indicator`, shown dashed). (*cols. 1–3*) pure, partial, full. (*col. 4*) Time-aggregated bar chart for pure, partial, and full effects.

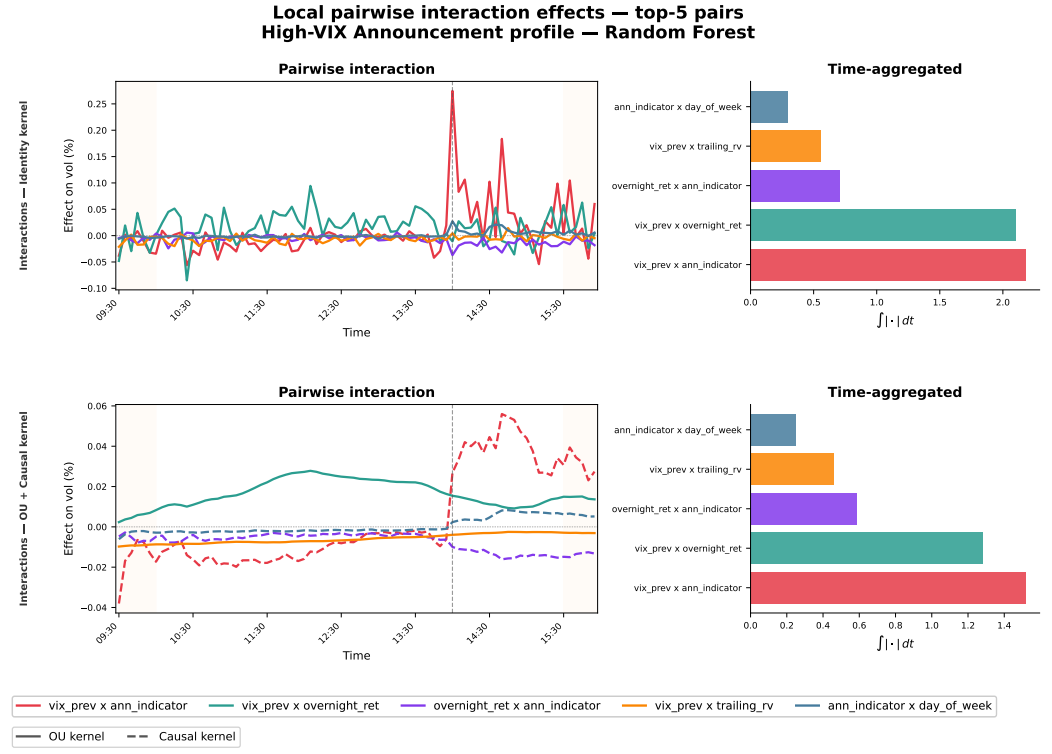


Figure 12: **Local pairwise interaction effects — top-5 pairs.** Same instance as Fig. 11. Top-5 pairs are ranked by time-integrated OU-kernel importance. (*row 1*) Identity kernel (pointwise Shapley interactions); (*row 2*) Feature-specific kernel (causal for pairs involving `ann_indicator`, shown dashed; OU otherwise). (*col. 1*) Interaction trajectories over the trading day; (*col. 2*) time-aggregated bar chart. The `vix_prev` × `ann_indicator` pair is the strongest interaction under both kernels.

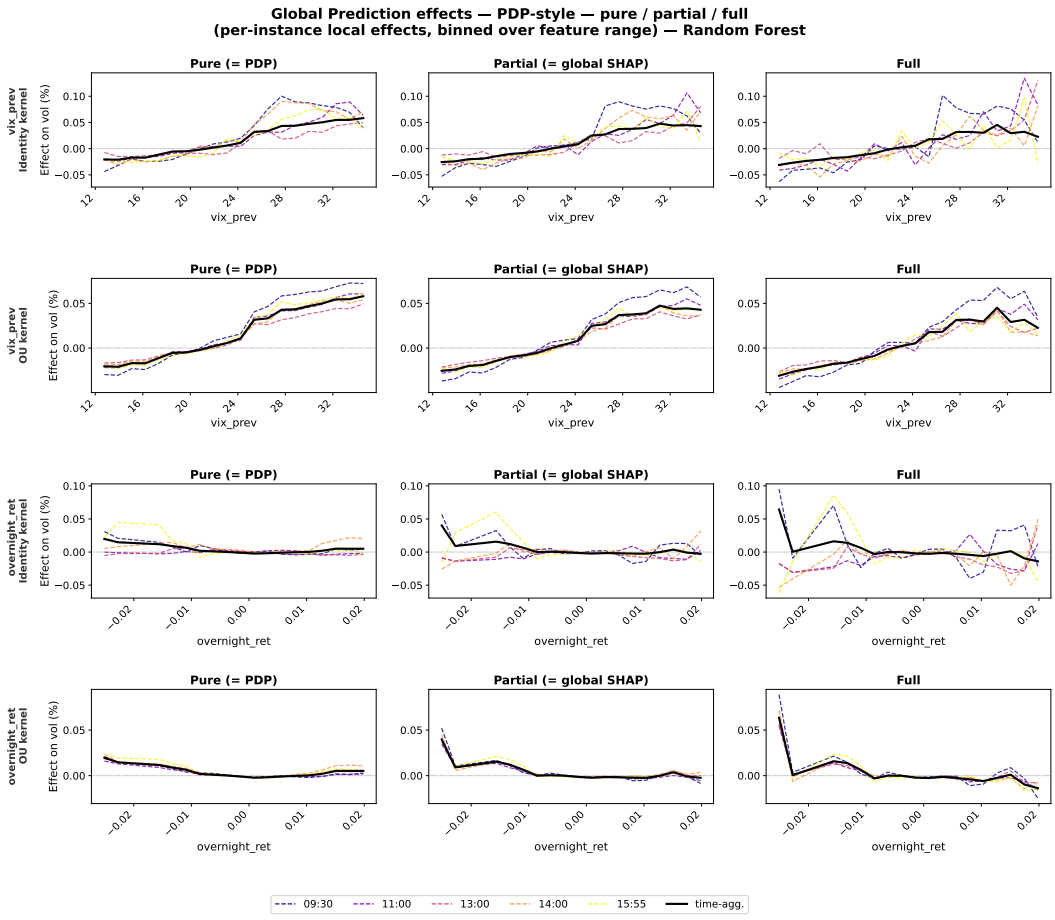


Figure 13: **Global prediction effects — PDP-style.** Per-instance pure, partial, and full local prediction effects, computed for $N_{\mathrm{PDP}} = 120$ profiles drawn by stratified sampling from the training set, are binned along the feature axis ($N_{\mathrm{bins}} = 20$) for the two highest-importance features by time-aggregated OU-kernel global prediction behavior (`vix_prev` and `overnight_ret`); (*rows 1–2*) `vix_prev` under the identity (*row 1*) and OU (*row 2*) kernels; (*rows 3–4*) `overnight_ret` under identity (*row 3*) and OU (*row 4*) kernels. Coloured dashed lines show the mean effect at five selected time points (09:30, 11:00, 13:00, 14:00, 15:55); the solid black line is the time-aggregated mean.

D.4 Electricity Demand Comparison

This appendix provides full experimental details and supplementary figures for the energy demand case study summarized in the main text (Sec. 5.2, Fig. 4).

Data. We use the UCI Individual Household Electric Power Consumption dataset (IHEPC) (Dua & Graff, 2019), which contains measurements of global active power at one-minute resolution for a single French household. We aggregate to hourly means, retaining only days with all $T = 24$ hours observed, yielding approximately $1{,}400$ complete days. The target trajectory is the hourly mean active power (kW). Additionally, we use the GB National Electricity System Operator historic demand dataset (NESO) (National Energy System Operator, 2024), which contains half-hourly settlement period demand figures (MW) for Great Britain. We use five years of data (2018–2022), retaining only days with all $T = 48$ half-hourly periods observed, yielding approximately $1{,}800$ complete days. The target trajectory is the half-hourly national demand in megawatts (MW), using the ND (national demand) column, falling back to TSD (transmission system demand) if unavailable. For both datasets the cross-day mean trajectory (the diurnal pattern) is estimated on the training split and subtracted from all splits, so the model is trained on day-to-day deviations from the typical daily profile.

Features. Six features are constructed for IHEPC and seven for NESO, all from information available at the start of the day:

Feature	Description	IHEPC	NESO
`day_of_week`	Integer $\in \{0, \ldots, 6\}$	✓	✓
`is_weekend`	Binary: 1 if Saturday or Sunday	✓	✓
`month`	Integer $\in \{1, \ldots, 12\}$	✓	✓
`season`	Integer $\in \{1, \ldots, 4\}$ (meteorological)	✓	✓
`lag_daily_mean`	Mean power/demand of the prior day	✓	✓
`lag_morning`	Mean power/demand during prior-day AM peak	✓	✓
`lag_evening`	Mean power/demand during prior-day PM peak		✓

The AM and PM peak windows are defined as 06:00–09:00 and 17:00–21:00 for IHEPC (hours 6–9 and 17–21) and 06:00–09:00 and 17:00–20:30 for NESO (half-hourly periods 12–18 and 34–41). Season is derived from month: winter $= \{12, 1, 2\}$, spring $= \{3, 4, 5\}$, summer $= \{6, 7, 8\}$, autumn $= \{9, 10, 11\}$.

Model. Both datasets use a chronological 80/20 train/test split.

Dataset	Split	# Days	%
IHEPC	Train	$\approx 1{,}120$	80%
	Test	≈ 280	20%
NESO	Train	$\approx 1{,}440$	80%
	Test	≈ 360	20%

Both datasets use the same multivariate random forest architecture as described in App. D.3, with identical hyperparameters (300 trees, `max_features="sqrt"`, `random_state=42`, `n_jobs=-1`). Each model predicts the full diurnal-adjusted trajectory jointly. The model achieves trajectory-level $R^2 = 0.191$ for IHEPC and $R^2 = 0.738$ for NESO, computed as $1 - \mathrm{SS}_{\mathrm{res}}/\mathrm{SS}_{\mathrm{tot}}$ over all test-split predictions jointly. The large difference reflects the fundamentally different nature of the two settings: NESO national demand is strongly driven by calendar and weather seasonality, which is well-captured by the available features, whereas single-household IHEPC demand is far more idiosyncratic, with individual behavior and appliance usage introducing substantial unexplained variability that no set of day-level features can capture. Both values are sufficient to produce stable and interpretable attributions.

Cooperative game set-up. The set-up follows App. D.3 in all respects except the kernel choice. The local prediction input behavior uses $n_{\mathrm{sample}} = 150$ background draws per coalition; PDP plots aggregate per-instance local effects over $N_{\mathrm{PDP}} = 120$ profiles drawn by stratified sampling from the training set. Global sensitivity and risk input behaviors are estimated by nested Monte Carlo with $n_{\mathrm{outer}} = n_{\mathrm{inner}} = 100$, using the same definitions as in App. D.3. With $p = 6$ (IHEPC) and $p = 7$ (NESO) features, there are $2^6 = 64$ and $2^7 = 128$ coalitions per game instance respectively.

Kernels. Two kernel choices are compared throughout.

- **Identity kernel** $K = I$. Treats each time step independently; corresponds to the pointwise baselines (PD/TsSHAP, Sobol, PFI/SAGE, depending on $\mathcal{B}_{\text{inp}}$).
- **Correlation kernel.** The empirical Pearson correlation matrix of the raw (non-diurnal-adjusted) training trajectories,

$$K_{st} = \frac{\mathrm{Cov}(Y_s, Y_t)}{\sqrt{\mathrm{Var}(Y_s)\,\mathrm{Var}(Y_t)}}, \qquad s, t \in [T],$$

clipped to $[-1, 1]$. This kernel encodes the empirical temporal dependence of demand: strongly correlated time steps receive similar attributions, so that a feature affecting the entire day is credited for its full coherent contribution rather than appearing as isolated hourly effects. Both kernels are row-normalized before application.

Profile selection. For the local explanations two profiles are studied, one per dataset. The IHEPC typical weekday profile is the median instance among 814 days satisfying `is_weekend` $= 0$ and `day_of_week` $\in \{1, 2, 3, 4\}$, with feature values `day_of_week` $= 4$ (Friday), `month` $= 11$ (November), `season` $= 4$ (autumn), `lag_daily_mean` $= 1.640$ kW, and `lag_morning` $= 2.324$ kW. The NESO winter weekday profile is the median instance among 322 days satisfying `is_weekend` $= 0$ and `season` $= 1$ (winter), with feature values `day_of_week` $= 0$ (Monday), `month` $= 12$ (December), `lag_daily_mean` $= 33{,}431$ MW, `lag_morning` $= 29{,}082$ MW, and `lag_evening` $= 41{,}313$ MW.

Supplementary figures. Figs. 14–20 show the full set of results. Color coding: `day_of_week` (blue), `is_weekend` (orange), `month` (green), `season` (red), `lag_daily_mean` (purple), `lag_morning` (brown), `lag_evening` (pink, NESO only). Solid lines correspond to the identity kernel; dashed lines to the correlation kernel. Blue and red shading marks AM and PM peak windows.

Figs. 14–15 show global sensitivity and risk effects under both kernels. For IHEPC, `lag_daily_mean`, `lag_morning`, and `month` dominate both sensitivity and risk across all effect types. Under the identity kernel their effects show separate morning and evening peaks, while the correlation kernel merges these into a single elevated region spanning both peaks, reflecting the block-diagonal covariance structure. Time-aggregated pure, partial and full time-aggregated effects diverge substantially, suggesting that forecast variance and error are sensitive to feature interactions. For NESO, `month` and `season` dominate, with substantially larger magnitudes than IHEPC consistent with national-level demand being driven by weather and calendar seasonality. Under the identity kernel effects vary across the day, but the correlation kernel collapses them to near-flat trajectories, reflecting the near-uniform covariance structure of the NESO dataset. The gap between time-aggregated pure and full effects for both risk and sensitivity is large for `season` and `month` in particular, indicating strong, focused higher-order interactions for these two features in particular at the national grid level, while other features participate less in interactions.

Sensitivity network plots. Fig. 16 summarizes the global sensitivity input behavior as a network: node size encodes the time-aggregated attribution of each feature, edge width encodes pairwise interaction strength, and edge color encodes the sign of the underlying Möbius coefficient (teal for positive, red for negative). Three columns show the pure, partial, and full effects respectively. A salient pattern is that negative edges (red) appear almost exclusively in the pure-effect column and largely vanish in the partial and full columns. This is expected. Pure effects are raw Möbius coefficients, which under exact Sobol orthogonality (independent features) equal the non-negative variance contributions, but pick up signed contributions whenever features are empirically dependent or when sampling noise is present. Partial effects (Shapley sensitivity) and full effects (total Sobol) are non-negative weighted sums of Möbius coefficients at all coalitions containing the feature in question, and these aggregations average out the signed pure-effect components. In our setting, the visible negative edges in the IHEPC pure column connect calendar features to each other and lag features, and in the NESO pure column they connect lag features to one another and season and month, both consistent with the known empirical correlation between these features.

Figs. 17–18 show local prediction effects for the selected profiles. For the IHEPC typical weekday, several features contribute comparably (`month`, `lag_daily_mean`, `day_of_week`, `lag_morning`), with no single dominant driver. Under the identity kernel, effects are concentrated

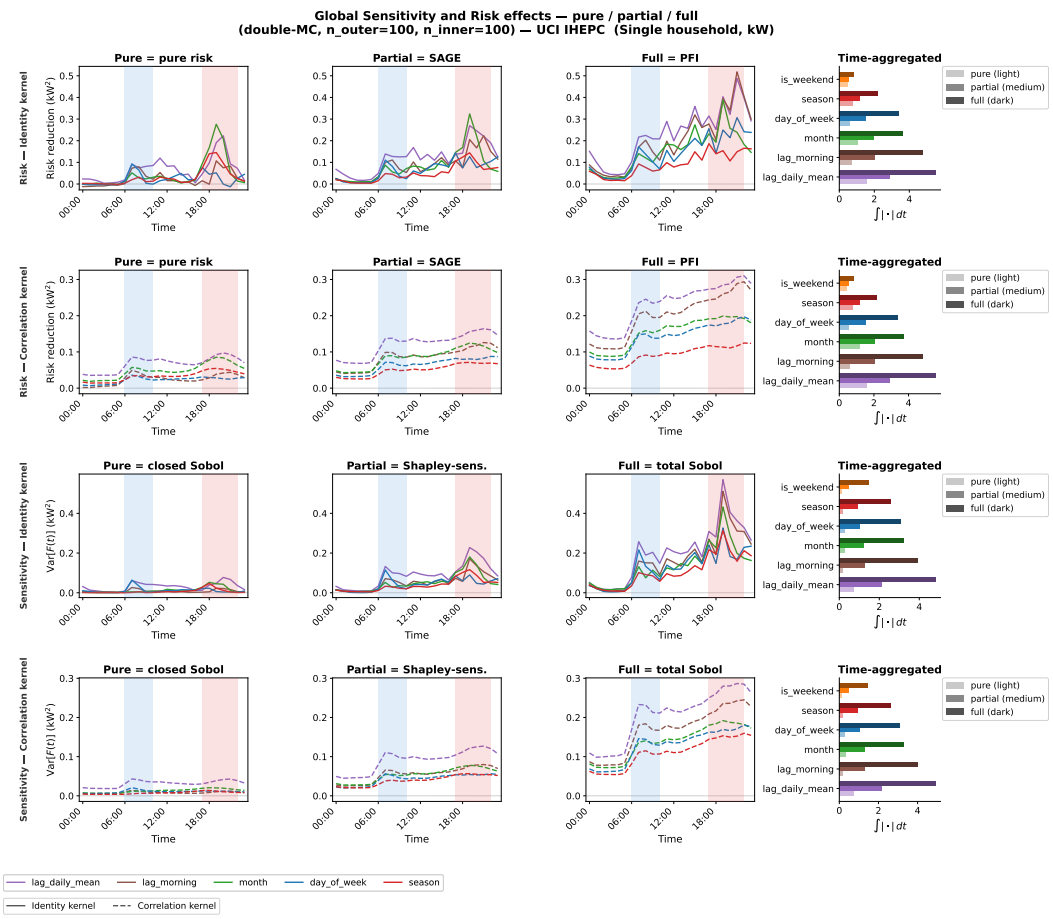


Figure 14: **Global sensitivity and risk effects — IHEPC (single household, kW).** Effects are computed by nested Monte Carlo ($n_{\text{outer}} = n_{\text{inner}} = 100$) directly from the global sensitivity and risk games. (*rows 1–2*) Risk game (loss reduction, kW^2) under the identity kernel (*row 1*) and correlation kernel (*row 2*); (*rows 3–4*) sensitivity game ($\text{Var}[F(t)]$, kW^2) under the identity (*row 3*) and correlation (*row 4*) kernels. (*cols. 1–3*) Pure (closed Sobol / pure risk), partial (Shapley-sensitivity / SAGE), full (total Sobol / PFI); (*col. 4*) time-aggregated bar chart. Blue and red shading marks AM and PM peak windows.

in the morning and evening peak windows, while the correlation kernel smooths these into broader positive contributions throughout the day with attenuated peaks. For the NESO winter weekday, contributions are similarly distributed across multiple features (notably `season`, `day_of_week`, `lag_evening`, and `month`), with the correlation kernel producing roughly constant positive effects across the day, consistent with winter demand being uniformly elevated relative to the annual mean. The correlation kernel substantially reduces the apparent effect magnitude (note the compressed y-axis), reflecting the fact that coherent daily shifts are re-expressed as a single weighted attribution rather than summed over independent time steps.

Figs. 19–20 show local pairwise interaction effects. For IHEPC the depicted pairs, which are the top-5 pairs by time-aggregated importance, frequently involve lag features interacting with calendar features (e.g., `month` × `lag_morning`, `day_of_week` × `lag_morning`), suggesting that the strength of autoregressive momentum depends on the time of year or day. Under the identity kernel interaction trajectories are noisy, but exhibit morning and evening peaks; the correlation kernel smooths these into coherent daily profiles. For NESO the `month` × `season` pair dominates by a wide margin, reflecting the near-collinearity of these two features and their joint encoding of seasonal demand patterns. Under the correlation kernel all interaction trajectories collapse to flat lines, consistent with the uniform covariance structure.

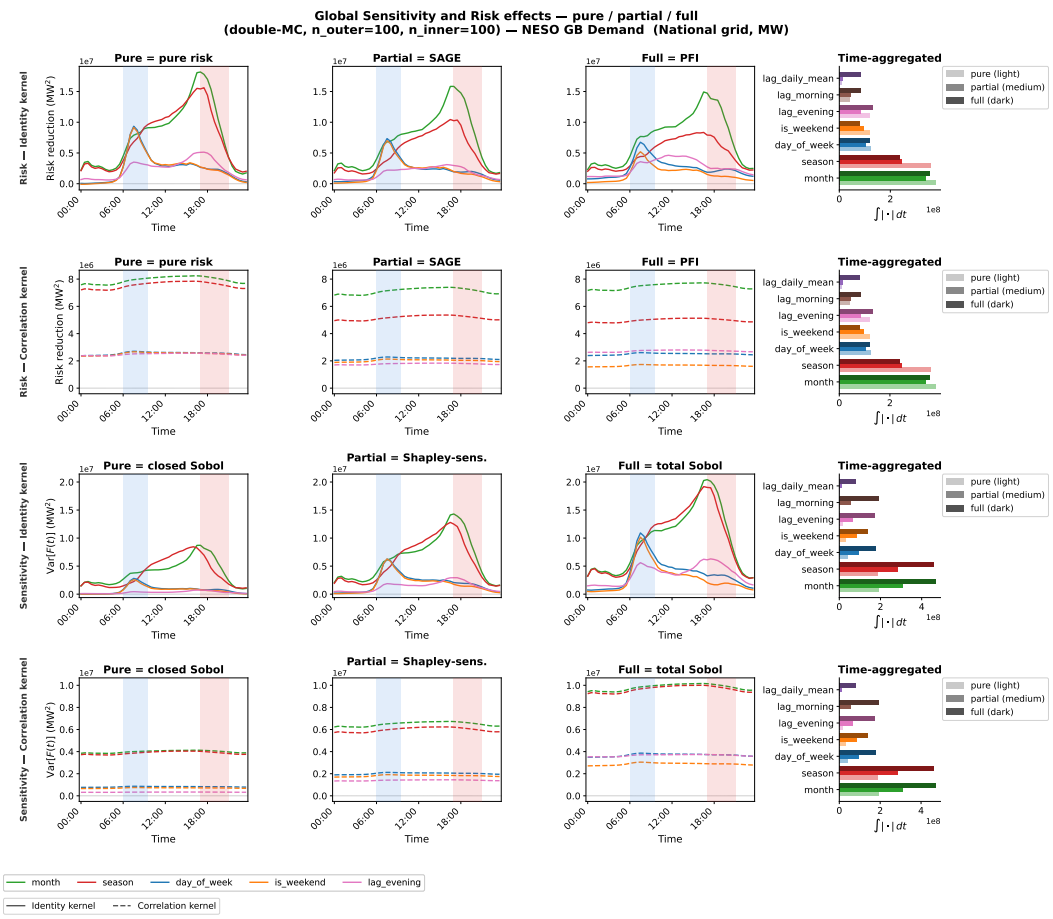


Figure 15: **Global sensitivity and risk effects — NESO (national grid, MW).** Effects are computed by nested Monte Carlo ($n_{\text{outer}} = n_{\text{inner}} = 100$) directly from the global sensitivity and risk games. (*rows 1–2*) Risk game (loss reduction, $\text{MW}^2 \times 10^7$) under the identity kernel (*row 1*) and correlation kernel (*row 2*); (*rows 3–4*) sensitivity game ($\text{Var}[F(t)]$, $\text{MW}^2 \times 10^7$) under the identity (*row 3*) and correlation (*row 4*) kernels. (*Cols. 1–3*) Pure (closed Sobol / pure risk), partial (Shapley-sensitivity / SAGE), full (total Sobol / PFI); *col. 4*: time-aggregated bar chart. Blue and red shading marks AM and PM peak windows.

Figs. 21–22 show global prediction effects in PDP style. For IHEPC the top-2 features by identity-kernel SHAP importance are `lag_daily_mean` and `month`, with `lag_daily_mean` showing a clear monotone relationship: higher prior demand predicts higher demand deviations. The correlation kernel preserves this monotone shape but tightens the effect and time point spread, indicating that the relationship is more coherent across the day rather than time-step-specific. `month` exhibits a U-shaped curve, capturing the January peak, summer trough, and second winter peak. For NESO the top-2 features are `month` and `lag_evening`. `month` exhibits a U-shaped curve again. The correlation kernel suppresses time point variability and cleanly reveals these seasonal patterns for both features.

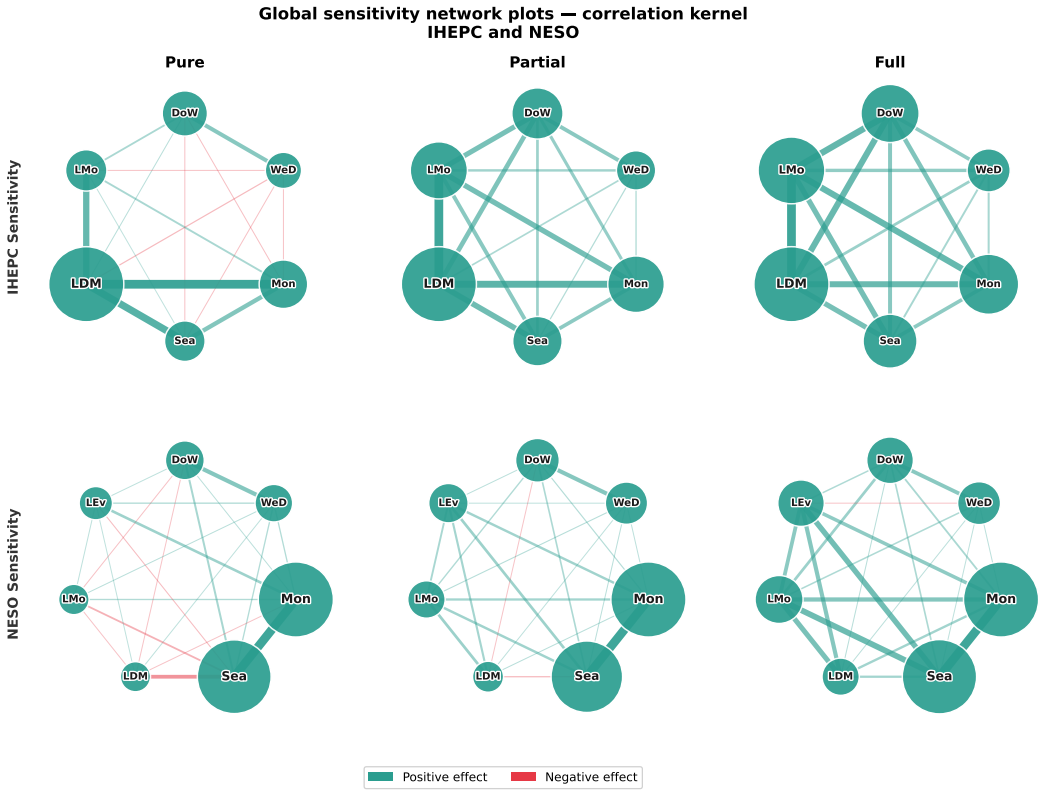


Figure 16: **Global sensitivity network plots — correlation kernel.** (*row 1*) IHEPC; (*row 2*) NESO; (*cols. 1–3*) pure, partial, and full effects. Node size encodes time-aggregated feature importance; edge width encodes pairwise interaction strength; edge color encodes the sign of the interaction (teal for positive, red for negative). Negative edges are concentrated in the pure-effect column and largely vanish after Shapley (partial) or total-Sobol (full) aggregation, which are non-negative weighted sums of Möbius coefficients. Feature abbreviations: `DoW` = `day_of_week`, `WeD` = `is_weekend`, `Mon` = `month`, `Sea` = season, `LDM` = `lag_daily_mean`, `LMo` = `lag_morning`, `LEv` = `lag_evening` (NESO only).

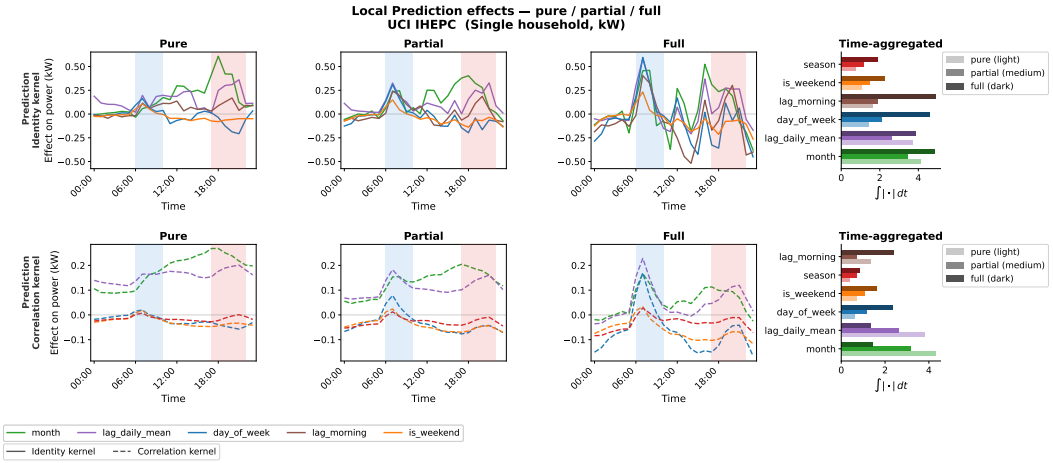


Figure 17: **Local prediction effects — IHEPC typical weekday.** Instance: `day_of_week` $= 4$, `month` $= 11$, `lag_daily_mean` $= 1.640\,\text{kW}$, selected as the median among 814 days with `is_weekend` $= 0$ and `day_of_week` $\in \{1, 2, 3, 4\}$. (*row 1*) Identity kernel; (*row 2*) correlation kernel; (*cols. 1–3*) pure, partial, full; (*col. 4*) time-aggregated bar chart.

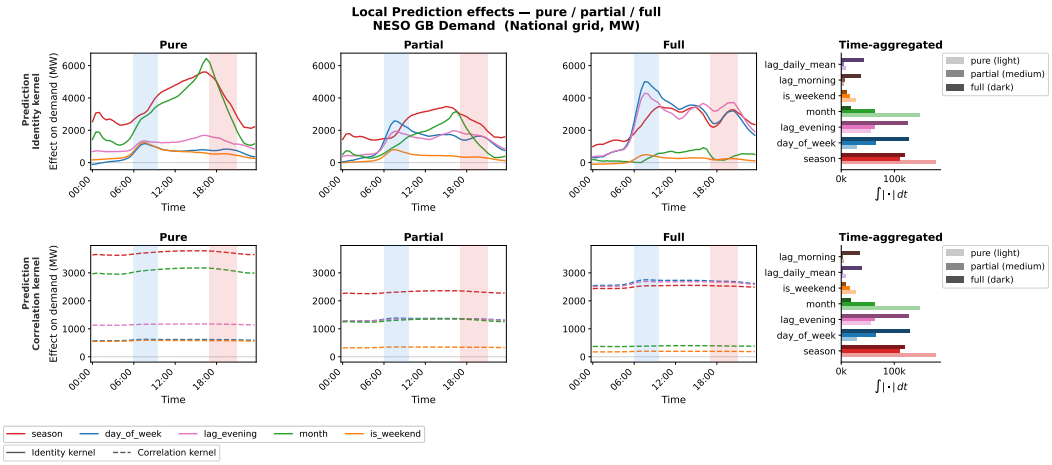


Figure 18: **Local prediction effects — NESO winter weekday.** Instance: `day_of_week` $= 0$, `month` $= 12$, `lag_daily_mean` $= 33{,}431$ MW, selected as the median among 322 days with `is_weekend` $= 0$ and `season` $= 1$ (winter). (*row 1*) Identity kernel; (*row 2*) correlation kernel. (*cols. 1–3*) Pure, partial, full; (*col. 4*) time-aggregated bar chart.

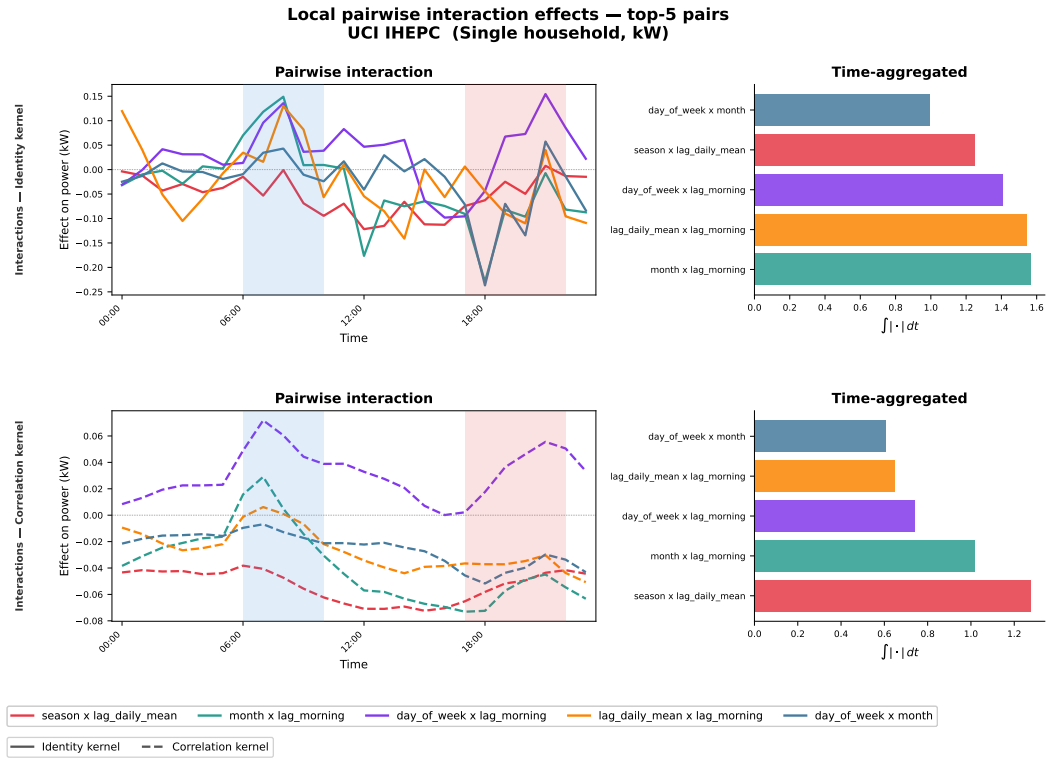


Figure 19: **Local pairwise interaction effects — top-5 pairs.** Instance: `day_of_week` $= 4$, `month` $= 11$, `lag_daily_mean` $= 1.640$ kW, selected as the median among 814 days with `is_weekend` $= 0$ and `day_of_week` $\in \{1, 2, 3, 4\}$. Top-5 pairs are selected by time-integrated correlation-kernel importance. (*row 1*) Identity kernel; (*row 2*) correlation kernel (dashed lines). (*col. 1*) Interaction trajectories over the 24-hour day; (*col. 2*) time-aggregated bar chart. Blue and red shading marks AM and PM peak windows.

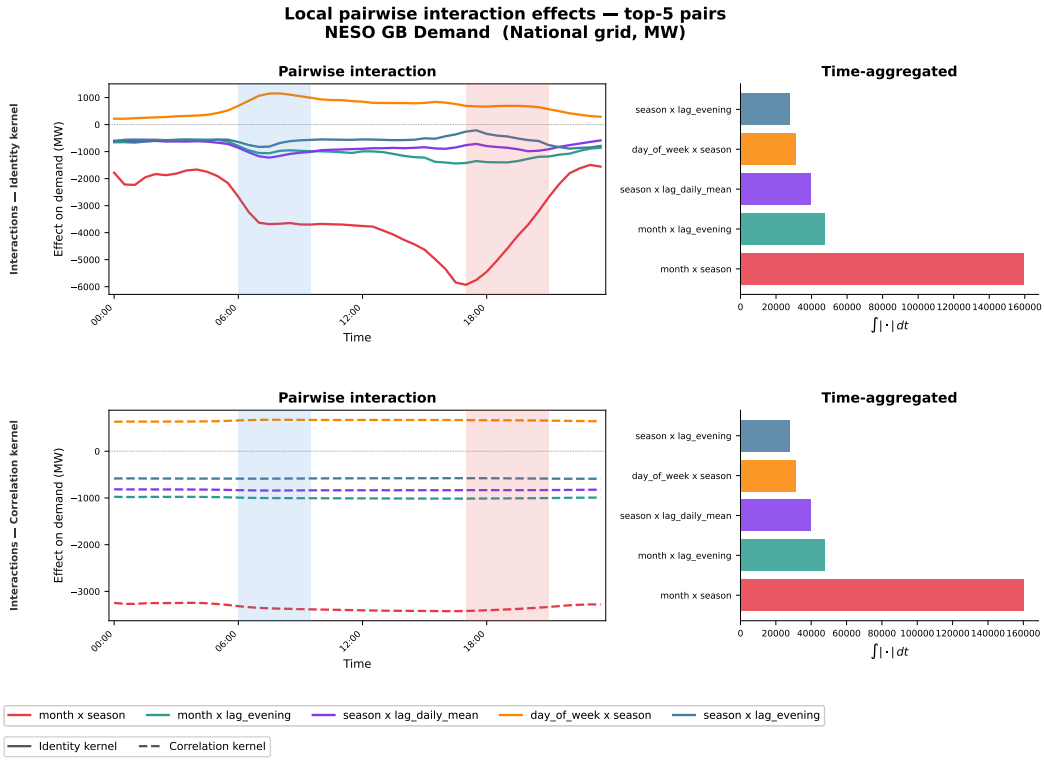


Figure 20: **Local pairwise interaction effects — top-5 pairs.** Instance: `day_of_week` = 0, `month` = 12, `lag_daily_mean` = 33,431 MW, selected as the median among 322 days with `is_weekend` = 0 and `season` = 1 (winter). Top-5 pairs are ranked by time-integrated correlation-kernel importance. (*row 1*) Identity kernel; (*row 2*) correlation kernel (dashed lines). (*col. 1*) Interaction trajectories over the 48 half-hourly periods; (*col. 2*) time-aggregated bar chart. The `month` × `season` pair dominates by a wide margin under both kernels. Blue and red shading marks AM and PM peak windows.

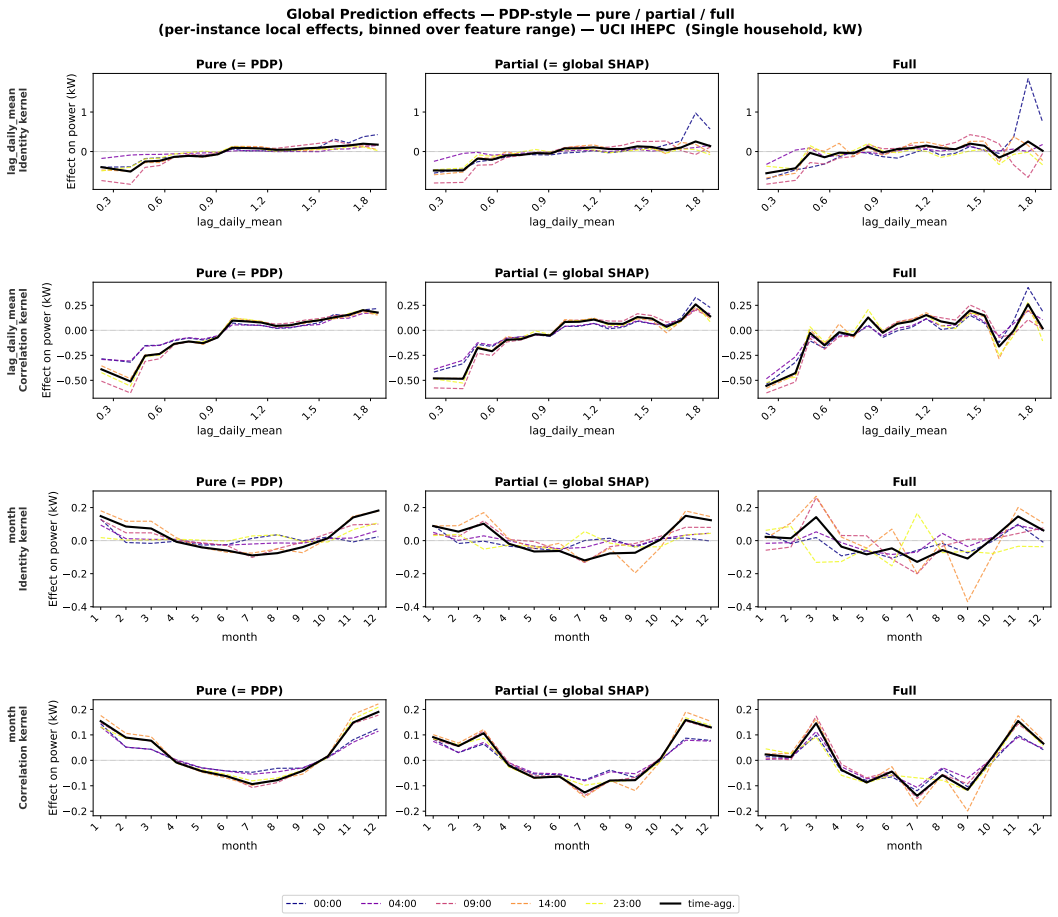


Figure 21: **Global prediction effects — PDP-style — IHEPC.** Per-instance pure, partial, and full local prediction effects, computed for $N_{\mathrm{PDP}} = 120$ profiles drawn by stratified sampling from the training set, are binned over the feature range ($N_{\mathrm{bins}} = 20$) for the two highest-importance features by identity-kernel SHAP importance. (*rows 1–2*) `lag_daily_mean` under the identity (*row 1*) and correlation (*row 2*) kernels; (*rows 3–4*) `month` under the identity (*row 3*) and correlation (*row 4*) kernels. Coloured dashed lines show mean effects at five selected time points (00:00, 04:00, 09:00, 14:00, 23:00); the solid black line is the time-aggregated mean.

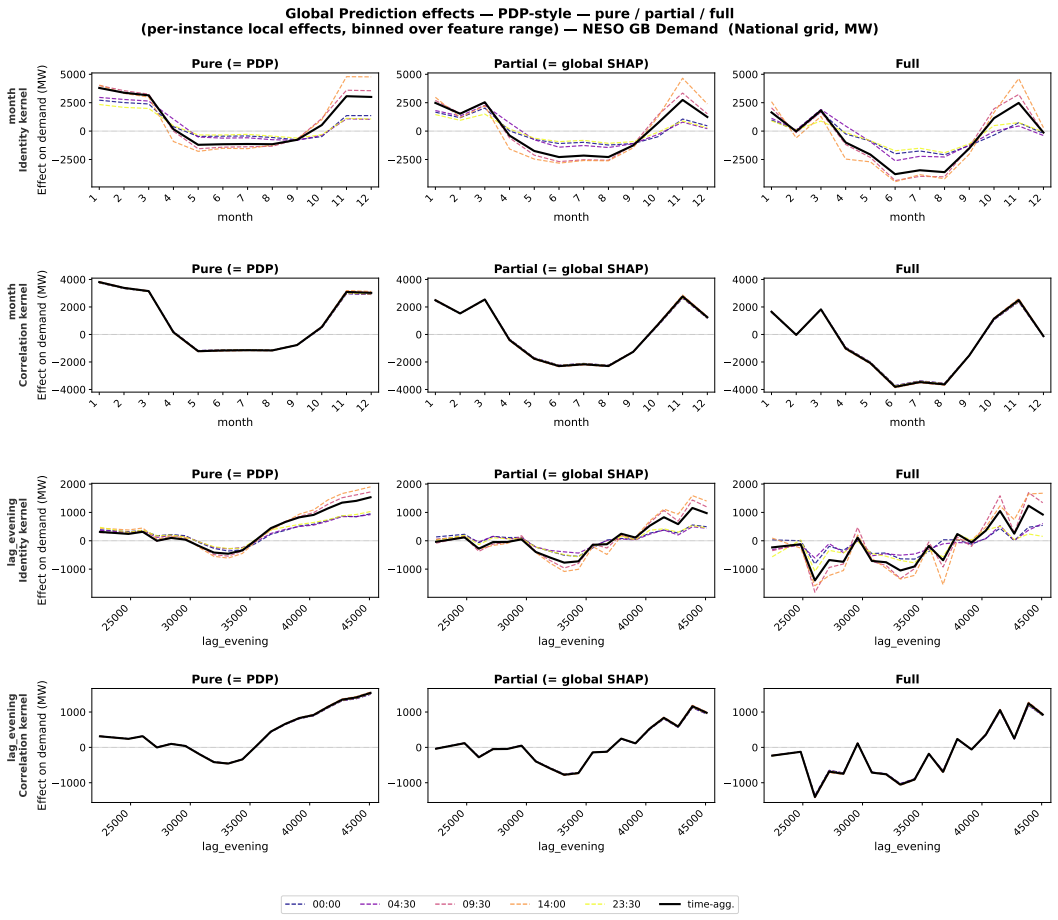


Figure 22: **Global prediction effects — PDP-style — pure / partial / full - NESO.** Per-instance pure, partial, and full local prediction effects, computed for $N_{\mathrm{PDP}} = 120$ profiles drawn by stratified sampling from the training set, are binned over the feature range ($N_{\mathrm{bins}} = 20$) for the two highest-importance features by identity-kernel SHAP importance; (*rows 1–2*) `month` under the identity (*row 1*) and correlation (*row 2*) kernels; (*rows 3–4*) `lag_evening` under the identity (*row 3*) and correlation (*row 4*) kernels. Coloured dashed lines show mean effects at five selected time points (00:00, 04:00, 09:00, 14:00, 23:00); the solid black line is the time-aggregated mean.

D.5 Quantitative comparison: temporal localization

Whether an explanation respects the temporal order of an event is measurable without committing to a single notion of faithfulness: if a feature's influence cannot begin before a known time, the share of its attribution placed earlier is a criterion any method can be scored against, whatever quantity it computes.

We compare them on a criterion that does not require a single notion of faithfulness across methods: given a feature whose influence cannot begin before a known time t_0, we measure the share of its attribution mass placed before t_0,

$$\rho(\phi, t_0) \;=\; \frac{\int_{t<t_0} |\phi(t)|\,\mathrm{d}t}{\int_{\mathcal{T}} |\phi(t)|\,\mathrm{d}t}.$$

A kernel that respects temporal order should keep ρ small. The anti-causal kernel (same functional form, same smoothing amount, reversed direction) serves as a control throughout.

On SPY we use the `ann_indicator` feature and the FOMC statement release at 14:00 ET, which is the subset of announcement days whose release falls inside the trading session; the 52 remaining announcement days carry an 08:30 ET print and therefore have no intraday pre-event window. This yields 19 days, including the instance studied in Sec. 5.2. Two controls accompany every comparison: a *placebo cutoff* at 11:00 ET, where nothing is scheduled, and the *control feature* `vix_prev`, which has no event structure.

Table 6 reports the aggregate. Under the pure interaction operator, `ann_indicator` is binary and takes the same value on all announcement days, so its pure effect is identical across days by construction; the reported value is therefore not instance-specific, and no dispersion is reported. Under the partial (Shapley) operator, interaction mass active before the release is redistributed onto the feature, which both reduces the effect and makes it vary across days.

Table 6: Temporal localization on SPY, averaged over the 19 intraday release days. *Reduction* is the paired per-day quantity $\rho(\text{identity}) - \rho(\text{causal})$, so positive means less pre-event mass; *excess* subtracts the reduction at the placebo cutoff, isolating the part not explained by the generic forward shift of a lower-triangular kernel. Causal and anti-causal length-scales are matched per day and feature to the OU smoothing amount ($\ell = 8$ bars). Values are mean $\pm$ sd across days; the pure effect of `ann_indicator` is constant across days (see text).

		pre-event share ρ					
Feature	Cutoff	identity	+ OU	causal	anti-causal	reduction	excess
Pure effects							
`ann_indicator`	release	0.260	0.253	**0.145**	0.361	$+0.115$	$+0.116$
`ann_indicator`	placebo	0.104	0.056	0.106	0.037	-0.001	—
`vix_prev`	release	0.667	0.676	0.730	0.646	-0.063	-0.005
`vix_prev`	placebo	0.290	0.283	0.347	0.238	-0.058	—
Partial effects							
`ann_indicator`	release	0.296	0.253	**0.211**	0.380	$+0.085$	$+0.045$
`ann_indicator`	placebo	0.117	0.052	0.077	0.049	$+0.040$	—
`vix_prev`	release	0.546	0.599	0.630	0.593	-0.084	-0.027
`vix_prev`	placebo	0.247	0.237	0.304	0.198	-0.057	—

Three observations. First, for the event feature the causal kernel reduces the pre-event share while the symmetric OU kernel barely moves it ($0.260 \to 0.253$ for pure effects), consistent with the claim that symmetric post-processing cannot remove pre-event mass. Second, the reduction is specific to the release time: at the placebo cutoff it is -0.001 (pure) and $+0.040$ (partial), so the excess column, not the raw reduction, is the localization effect. Third, and most directly against the concern that a lower-triangular kernel trivially lowers ρ, the causal kernel *increases* the pre-event share of the control feature `vix_prev` on 19/19 days (pure) and 17/19 days (partial). The direction of the effect therefore depends on whether the feature's influence begins at the cutoff or spans the session, which is exactly what the criterion is intended to detect.

Under the partial operator the reduction is positive on 15 of 19 days (sign test $p = 0.019$) and the excess over placebo on 16 of 19 ($p = 0.004$). The four exceptions are 2022-06-15, 2022-07-13, 2022-09-21 and 2023-03-22. Per-day values are given in Table 7.

Table 7: Per-day pre-event share of `ann_indicator` at the release bar, partial (Shapley) effects. The pure effect is identical on all days (Table 6) and is therefore not repeated. The last column reports the paper's nominal causal kernel ($\ell = 8$ bars) alongside the smoothing-matched one.

Date	identity	+ OU	causal (matched)	anti-causal	causal ℓ=8
2022-01-26	0.326	0.237	0.238	0.351	0.241
2022-03-16	0.337	0.266	0.242	0.381	0.244
2022-05-04	0.294	0.273	0.213	0.393	0.216
2022-06-15	0.298	0.304	0.301	0.383	0.306
2022-07-13	0.316	0.337	0.329	0.455	0.327
2022-07-27	0.326	0.253	0.272	0.387	0.274
2022-09-21	0.246	0.295	0.250	0.429	0.245
2022-11-02	0.280	0.253	0.155	0.404	0.154
2022-12-14	0.272	0.209	0.157	0.369	0.155
2023-02-01	0.289	0.178	0.092	0.333	0.092
2023-03-22	0.441	0.459	0.469	0.501	0.451
2023-05-03	0.301	0.175	0.141	0.329	0.141
2023-06-14	0.243	0.228	0.170	0.356	0.168
2023-07-26	0.262	0.248	0.238	0.340	0.236
2023-09-20	0.284	0.221	0.189	0.344	0.186
2023-11-01	0.381	0.218	0.205	0.364	0.206
2023-12-13	0.244	0.233	0.175	0.385	0.173
2024-01-31	0.259	0.175	0.099	0.321	0.099
2024-03-20	0.217	0.243	0.078	0.384	0.078

D.6 Relation to post-hoc smoothing of pointwise attributions

An analyst faced with noisy pointwise attributions will naturally consider computing them first and smoothing them afterwards, for example by a sliding-window average over neighboring output locations. This is the most direct alternative to our framework under prediction input behavior, and it is worth stating precisely how the two relate.

Under prediction input behavior the two coincide. Applying the output kernel to the game values before the interaction operator, as our framework does, and applying the same kernel to the resulting pointwise attributions afterwards give identical explanations. The reason is that the two operations act on different arguments and both are linear: the interaction operator is a fixed weighted sum over coalitions, whereas the kernel acts on the output domain, and the prediction input behavior is linear in the model output, so the order in which they are applied does not matter. We verified this on the SPY cooperative game of Sec. 5.2, across all kernels considered in App. D.5 and all three interaction operators: the two computation orders agree to within floating-point rounding.

The smoothed pointwise baseline is therefore not a competing method under prediction input behavior but a special case of our framework. In particular, the identity-kernel results in Sec. 5.2 are the pointwise methods of Sec. 4.3, and their smoothed variants are obtained by the corresponding kernel without any separate implementation.

This should not be read as the framework reducing to smoothing. The equivalence runs in one direction only: everything reachable by post-processing is reachable within the framework, but not conversely, and it fails in two distinct ways. Non-symmetric kernels, such as the causal kernel used for `ann_indicator`, redistribute attribution in one temporal direction and cannot be obtained from a symmetric smoothing operation at any bandwidth; this gap is what App. D.5 quantifies. And for the sensitivity and risk input behaviors the explained quantity is nonlinear in the model output, so the argument above no longer applies and kernel aggregation cannot be reproduced by post-processing pointwise explanations (cf. Remark 1). In those cases no analogous post-processing baseline exists at all.

D.7 COMPUTATIONAL COST

The dominant cost is inherited from the underlying cooperative game rather than introduced by the output behavior operator: the exact Möbius transform requires 2^p coalition evaluations, each averaging over n_{sample} background draws, identical to the pointwise methods our framework generalizes. Because our model predicts the full trajectory jointly, one coalition evaluation returns all T output locations at once, so the output dimension enters as a factor T in memory and in arithmetic per coalition rather than as additional model calls; a model that had to be queried per output location would instead incur a factor T there as well. Global input behaviors repeat this per explained instance. The identity kernel reduces none of these terms and no kernel increases them, and since the construction leaves the game unchanged, standard approximations for Shapley values and interactions apply directly (Štrumbelj & Kononenko, 2014; Fumagalli et al., 2023).

Two points follow. The exponential term arises only for the partial (Shapley) operator. First-order pure and full effects require only $\mathcal{O}(p)$ coalition evaluations and recover PD/closed-Sobol and PFI/total-Sobol, respectively. Thus, for kernel-aware explanations, all three levels of granularity remain available when exact Shapley computation is prohibitive. The cost grows with interaction order, requiring $\mathcal{O}(p^2)$ evaluations for all pairwise effects and $\mathcal{O}(p^k)$ for order k. Hence, it is the depth of the interaction structure being explained, rather than the kernel, that drives the computational expense. The kernel itself adds only a single $T \times T$ matrix product per Möbius coefficient ($2^6 \times 78^2$ for SPY, $2^7 \times 48^2$ for NESO), independent of p and negligible relative to the model evaluations; applied to cached game values, it also allows different kernels to be compared without additional model calls, as in the sweeps of App. D.5. Measured runtimes are reported in App. D.9.

D.8 DATA AVAILABILITY

Intraday five-minute OHLCV bar data for SPY were obtained from Polygon.io (`https://polygon.io`) covering January 2022 to April 2024. Access requires a Polygon.io subscription. The raw bars and the derived feature matrix therefore cannot be redistributed. Instead, the repository provides the precomputed cooperative-game caches from which all SPY figures can be regenerated without access to the raw data. The IHEPC dataset is publicly available from the UCI Machine Learning Repository (Dua & Graff, 2019). GB historic demand data are publicly available from the National Energy System Operator data portal (National Energy System Operator, 2024). All analysis code is available at GitHub.

D.9 COMPUTATIONAL DETAILS

Computing environment. All experiments were run on a 64-bit Linux platform running Ubuntu 22.04 LTS with two AMD EPYC Genoa 9534 64-core processors (128 cores, 256 threads total), 1.5 TB of RAM, and eight NVIDIA RTX 6000 Ada Generation GPUs (each with 48 GB memory). GPUs were not used: all model training and cooperative-game computations rely on CPU-only `scikit-learn` routines, parallelized over cores via `n_jobs=-1`.

Approximate wall-clock times. The synthetic ground-truth recovery experiment (Sec. D.1), comprising 30 Monte Carlo runs $\times 8$ sample sizes $\times 4$ model classes plus the oracle, with exact 2^p Möbius computation at each setting, completes in approximately 30–45 minutes. The SPY case study (Sec. D.3) completes in approximately 15 minutes for a first run, dominated by the global-game cache (30 instances $\times$ 3 game types). The IHEPC energy experiment (Sec. D.4) takes approximately 20 minutes for a first run, and the NESO experiment approximately 45 minutes; NESO is slower due to its larger coalition space ($2^7 = 128$ vs. $2^6 = 64$ for IHEPC) and its finer half-hourly resolution ($T = 48$ vs. $T = 24$).