*The Annals of Applied Statistics*
2023, Vol. 17, No. 2, 1260–1284
https://doi.org/10.1214/22-AOAS1668


# DETECTING DISTRIBUTIONAL DIFFERENCES IN LABELED SEQUENCE DATA WITH APPLICATION TO TROPICAL CYCLONE SATELLITE IMAGERY

BY TREY MCNEELY[1,a], GALEN VINCENT[1,b], KIMBERLY M. WOOD[2,d], RAFAEL IZBICKI[3,e] AND ANN B. LEE[1,c]

[1]*Department of Statistics and Data Science, Carnegie Mellon University,* [a]*treymcneely@gmail.com,* [b]*galenbvincent@gmail.com,* [c]*annlee@andrew.cmu.edu*

[2]*Department of Geosciences, Mississippi State University,* [d]*kimberly.wood@msstate.edu*

[3]*Department of Statistics, Federal University of São Carlos,* [e]*rafaelizbicki@gmail.com*

Our goal is to quantify whether, and if so how, spatiotemporal patterns in tropical cyclone (TC) satellite imagery signal an upcoming rapid intensity change event. To address this question, we propose a new nonparametric test of association between a time series of images and a series of binary event labels. We ask whether there is a difference in distribution between (dependent but identically distributed) 24-hour sequences of images preceding an event vs. a nonevent. By rewriting the statistical test as a regression problem, we leverage neural networks to infer modes of structural evolution of TC convection that are representative of the lead-up to rapid intensity change events. Dependencies between nearby sequences are handled by a bootstrap procedure that estimates the marginal distribution of the label series. We prove that type I error control is guaranteed as long as the distribution of the label series is well estimated which is made easier by the extensive historical data for binary TC event labels. We show empirical evidence that our proposed method identifies archetypes of infrared imagery associated with elevated rapid intensification risk, typically marked by deep or deepening core convection over time. Such results provide a foundation for improved forecasts of rapid intensification.

**1. Introduction.** A broad array of problems in the physical, environmental and biological sciences feature high-dimensional time series $\{\mathbf{X}_t\}_{t\geq 0}$, associated with binary "labels" $\{Y_t\}_{t\geq 0}$ indicating an event of interest. Examples include sequences of satellite or other remote sensing data paired with natural events like the occurrence of an earthquake, the rapid intensification of a hurricane or multivariate electroencephalographic (EEG) and magnetoencephalographic (MEG) data showing brain activity paired with physiological events like the occurrence of a stroke. Most research on this front concerns prediction of events (Luo et al. (2014)), measurement of event impact (Scharwächter and Müller (2020a, 2020b)) or detection of change points (Aminikhanghahi and Cook (2017), Evans and G'Sell (2020)) after events occur. Furthermore, joint analyses of time and event series often assume that the time series is univariate or model the relationship between multiple scalar quantities, as they change over time. There is a lack of theoretically and computationally sound methods for (nonparametric) association studies and statistical tests for dependent *and* high-dimensional sequence data.

This work is motivated by the need to identify spatiotemporal patterns in the convective evolution of tropical cyclone satellite imagery prior to a rapid intensity change; see "Motivating Application." Our immediate goal is not operational forecasting or prediction per se but rather gaining scientific insight into the spatiotemporal evolution $\mathbf{S}_{<t} =$

Received February 2022; revised July 2022.

*Key words and phrases.* Two-sample testing, high-dimensional time series, association studies, remote sensing, functional data, weather forecasting.

$\{\mathbf{X}_{t-T}, \mathbf{X}_{t-T+1}, \ldots, \mathbf{X}_t\}$ of convective structure or satellite imagery $\mathbf{X}_t$, leading up to a rapid intensity change event ($Y_t = 1$), for some lead time $T$, and identifying whether it differs in distribution from sequences $\mathbf{S}_{<t}$ that precede a nonevent ($Y_t = 0$).

From a statistical methodology standpoint, this problem amounts to a challenging two-sample testing problem for high-dimensional *dependent but identically distributed* (DID) data. From the observed time and event series, we extract labeled sequence data $\{(\mathbf{S}_{<t}, Y_t)\}_{t\geq 0}$, which we assume is a stationary process. That is, both $\mathbf{S}_{<t}$ and $Y_t$ are autocorrelated and dependent for different instances of time $t$. By the assumption of stationarity, the data $(\mathbf{S}_{<t}, Y_t)$ are identically distributed. Given historical data, we test whether the distributions of $\mathbf{S}_{<t}|Y_t = 1$ and $\mathbf{S}_{<t}|Y_t = 0$ are the same. The challenge is to construct an efficient test of (1) that is valid (controls type I error) for DID data and that applies to different types of sequence data (images, functions and sequentially observed data from multiple physical probes) with a minimum of assumptions. Finally, for many problems of applied interest, scientists want to know not just whether sequences preceding an event vs. nonevent are significantly different in distribution, but if so, also *how* the two distributions are different. That is, if the null hypothesis that the distributions of (stationary) sequences $\mathbf{S}|Y = 1$ and $\mathbf{S}|Y = 0$ are the same is rejected, the question is how to identify the patterns in the state space $\mathcal{S}$ of $\mathbf{S}$ that contributed to the rejection.[1] These patterns correspond to sequences $\mathbf{s} \in \mathcal{S}$ that are more or less likely to be associated with an event ($Y = 1$) than by chance.

1.1. *Motivating application*: *Tropical cyclones*. Tropical cyclones (TCs) are highly structured storms which rank among the deadliest and costliest natural disasters in the United States (Klotzbach et al. (2018)). Cases of rapid intensification (RI) and rapid weakening (RW) of such storms—defined for this work as a change in maximum wind speed of, at least, 25 knots within 24 hours, denoted by $Y = 1$—are notoriously difficult to forecast (Kaplan and DeMaria (2003), Kaplan, DeMaria and Knaff (2010), Kaplan et al. (2015), Wood and Ritchie (2015)). RI prediction has thus been the "highest-priority forecast challenge" identified by the National Hurricane Center (NHC) in the last decade (Gall et al. (2013)). RW events, while a lower priority than RI, are also associated with above-average forecast errors and are of great interest to meteorologists.

Models, such as SHIPS-RII (Kaplan et al. (2015)), have made great progress on skillfully forecasting RI events, but these approaches rely on scalar predictors (e.g., area-averaged vertical wind shear) at fixed points in time and thus neglect the evolving spatial structure of the TC; these structural changes often influence such events. To address this gap, meteorologists and forecasters seek interpretable patterns in the spatiotemporal structure of physically-relevant 2D fields that could indicate an elevated risk of RI. The first step in this search is to find interpretable temporally-evolving sequences of spatial structure $\mathbf{S}_{<t}$ that differ in distribution, depending on whether a TC is undergoing a rapid intensity change event ($Y_t = 1$) or not ($Y_t = 0$).

Our application examines one such 2D field: deep convection within the storm. Convection, or deep thunderstorm-like clouds, is a key component of the mechanism through which TCs extract energy from the ocean, meaning that convection, in both strength and distribution, should be closely related to storm intensity. We quantify convective structure using cloud-top temperature, as measured by infrared (IR) imagery from the Geostationary Operational Environmental Satellites (GOES). As convection strengthens, the cloud tops are pushed higher into the atmosphere where temperatures are lower. The temperature of the cloud top can, therefore, be used as a proxy for the strength of convection in the storm. We ask: "Do

[1]The stationarity assumption makes the state space $\mathcal{S}$ well defined and allows us to drop the subindex $< t$ in our notation.

24-hour sequences of convective structure, $\{\mathbf{S}_{<t}\}_{t\geq 0}$, contain information about upcoming intensity change, $\{Y_t\}_{t\geq 0}$? If so, how do the spatiotemporal patterns of rapidly changing TCs differ from that of not rapidly changing TCs?"

Both $\{\mathbf{S}_{<t}\}_{t\geq 0}$ and $\{Y_t\}_{t\geq 0}$ are highly dependent (autocorrelated) time series in this application: environmental fields, such as convection, change slowly, and rapid intensity change events are, by definition, extended periods of change (typically, 12–48 hours), meaning that successive measurements of these variables at short time intervals (e.g., three hours) will be highly dependent. It is particularly challenging to perform traditional two-sample tests in this DID setting due to the combination of low sample size (671 TCs between 2000–2020), high-dimensional image data and strong temporal dependence.

1.2. *Contribution and relevance.* Our contribution is twofold: (i) On the methodology side we present a new statistical framework for detecting arbitrary distributional differences in a high-dimensional setting with labeled sequence data. The proposed two-sample test is valid in a DID setting and provides local diagnostics as to the type of sequences $\mathbf{s} \in \mathcal{S}$ that contribute to rejection of $H_0$ in equation (2). (ii) On the applied side we utilize our proposed framework to identify and describe patterns of convective evolution in TCs prior to the onset of rapid intensity change.

Figure 1 shows a flow chart of our overall approach. As indicated by the vertical blue arrow to the left, we seek quantitative conclusions regarding how TC behavior relates to convective structure, as observed by GOES imagery and extracted functions. We cast the two-sample test as a prediction problem (Figure 1, right, for "Statistical Methods"). This allows us to leverage powerful prediction techniques, such as convolutional neural nets (CNNs), to gain scientific insight from high-dimensional functional or video data without a prior dimension reduction (arrow back to "TC behavior" and "Intensity Guidance"). To account for dependence in the labeled sequence $\{(\mathbf{S}_{<t}, Y_t)\}_{t\geq 0}$, we develop a bootstrap regression test (Algorithm 1) which yields a valid p-value accompanied by interpretable diagnostics.

Our bootstrap test is ideal for TC studies: the low number of unique TCs for which high-resolution satellite imagery is available prohibits efficient inference via classical blocking schemes, whereas our method can take advantage of the extensive historical record of event (label) sequences. Theorem 1 shows that the bootstrap test is valid as long as we can estimate

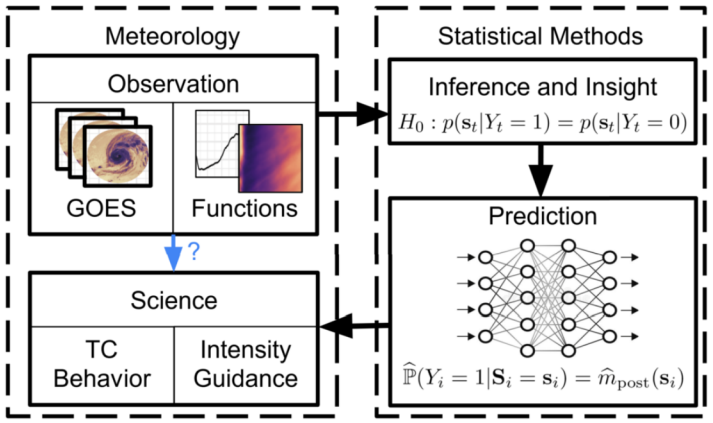



FIG. 1. *Leveraging prediction tools for statistical inference and scientific insight. In our application we seek relationships between high-dimensional TC observations and TC behavior. We pose this as a two-sample testing problem, but the high dimensionality and sequential correlation in our data make this hypothesis difficult to test. By rewriting the test in terms of a prediction problem, we are able to leverage powerful prediction methods to infer modes of structural evolution in TCs which indicate rapid intensity change.*

the distribution of the label sequences well; this is made easier by the fact that labels are binary.

Section 7 provides evidence that deep and/or deepening convection is a necessary precursor to RI but that other factors (e.g., low vertical shear) must be present. This elevated risk of RI due to deepening convection is often present *prior* to the onset of intensification, demonstrating value to forecasting pipelines. RW, meanwhile, is a more variable process and did not return significant results.

1.3. *Relation to other work.* Here we review some related works.

*Event impact and causal inference for time series.* Our problem setup is closest in flavor to association and causal inference studies for testing the relationship between a time and event series (Luo et al. (2014), Scharwächter and Müller (2020b)). The vast majority of these works assume univariate time series or test for each dimension in a multivariate time series separately (Candès et al. (2018)). The recent paper by Scharwächter and Müller (2020b) leverages a two-sample testing approach for high-dimensional data, as in this work, albeit to study how a discrete binary event history impacts a multivariate time series rather than how the evolution of a complex time series is associated with a later event. A key methodological difference is that their work handles the lack of independence by sampling data points that are far from each other, while our method allows for the use of all available data. This distinction is key for applications with limited data. Their proposed multiple testing procedure also does not control the false positive error rate exactly but heuristically.

*Two-sample tests in high dimension.* Recently, there has been a growing interest in nonparametric two-sample tests in high dimension. Popular machine learning-based approaches include classification accuracy tests (Kim et al. (2021)), kernel-based tests (Gretton et al. (2012)) and divergence-based density ratio tests (Kandasamy et al. (2015), Moon and Hero (2014)). We use the same regression test statistic (equation (4)) as Kim, Lee and Lei (2019) to allow for interpretable local diagnostics. However, Kim, Lee and Lei (2019) and the abovementioned two-sample testing papers only handle the standard independent and identically distributed (IID) data setting (Figure 2a), whereas the methods in this work apply to DID sequence data (Figure 2c).

Also related to our paper are modern tests of conditional independence between two random vectors $Y$ and $Z$, given a third random vector $X$ (see discussion in Section 8.2), and tests of the conditional mean and quantile dependence of $Y$ on $X$. Most high-dimensional research on this front (including model-X knockoffs; Berrett et al. (2020), Candès et al. (2018), Sesia, Sabatti and Candès (2019)) assumes IID data $(X_{i1}, \ldots, X_{ip}, Y_i) \sim F_{X,Y}$ for $i = 1, \ldots, n$ and also assumes that the distribution of $Y$, given $X$, depends on only a small fraction of the $p$ covariates. The latter sparsity assumption is reasonable for, for example, genome-wide association studies but not for remote sensing applications with image and functional data. Another key difference between our testing approach and so-called model-X methodologies, which also use machine learning algorithms to approximate the distribution of $Y$ given $X$ (Katsevich and Ramdas (2020)), is that model-X approaches estimate or make assumptions regarding the distribution of $X$ (or $X$ given $Z$), whereas we instead estimate the distribution of the response $Y$.

*Bootstrap for time series.* There exist many different types of bootstrap methods for dependent data; see, for example, Bühlmann (2002), Horowitz (2003), Kreiss and Paparoditis (2011) for a review. The goal is often to model the data distribution for parameter estimation, rather than as here to test for an association between a label series and a high-dimensional time series. Our framework also does not bootstrap the entire distribution of $\{(\mathbf{S}_{<t}, Y_t)\}_{t\geq 0}$ but only the distribution of labels $\{Y_t\}_{t\geq 0}$. For binary labels this is a much easier estimation problem than, for example, block-bootstrap of high-dimensional time series.

*TC analyses.* Many TC analysis tools incorporate information from 2D fields via scalar values such as area averages; for example, the operationally-used SHIPS and SHIPS-RII forecast schemes include scalars for the fraction of pixels with IR temperatures below $-30°$C within the 50–200-km annulus (DeMaria and Kaplan (1999), Kaplan et al. (2015)). Such approaches discard complex, time-evolving structure in the 2D fields. More recent analyses take spatial information into account by applying dimension reduction techniques, like functional principal component analysis (PCA), to the field, such as for TC eye formation forecasts in (Knaff and DeMaria (2017)). However, dimension reduction adds an extra layer of abstraction between TC structure and subsequent TC behavior and can reduce meteorologists' ability to interpret the information. The ORB framework (Organization, Radial structure, Bulk morphology) was introduced in McNeely et al. (2020) to summarize convective structure into a dictionary of functional features. The key objective was to utilize entire functions to quantify structure rather than thresholded feature statistics, thereby enabling richer descriptions of spatial structure while remaining interpretable. In this paper we leverage CNNs to model the relationship between TC intensity change and the temporal evolution of one such continuous ORB function—the radial profile of cloud-top temperatures, as observed by GOES-IR imagery (McNeely et al. (2019), McNeely et al. (2020), Sanabia, Barrett and Fine (2014)). Our bootstrap test then provides a powerful tool for directly assessing whether there is an association between RI or RW events and TC structure or the TC environment in terms of structural summaries (1D ORB functions) that are easily digestible to meteorologists.

1.4. *Outline.* We begin by defining the problem set up in Section 2, paying special attention to different dependence structures in $\{(\mathbf{S}_{<t}, Y_t)\}_{t\geq 0}$. In Section 3 we describe the TC data. In Section 4 we lay out the details of our bootstrap test for distributional differences in dependent sequence data. In Section 5 we provide theoretical justification for validity of the bootstrap test. In Section 6 we introduce a simulated toy example to empirically demonstrate the advantage of a Markov chain-based bootstrap test over traditional permutation testing. In Section 7 we apply our method to study the evolution of convective structure in TCs prior to a rapid intensity change. Finally, in Section 8 we discuss limitations and potential extensions of our method.

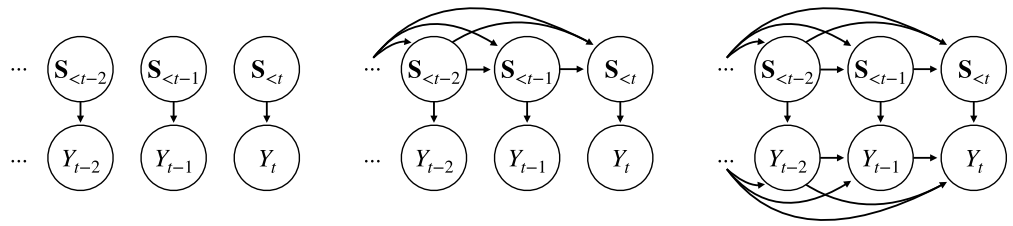
**2. Setup.** Our goal is to detect distributional differences in labeled sequence data $\{(\mathbf{S}_{<t}, Y_t)\}_{t\geq 0}$, where the "labels" $Y_t \in \{0, 1\}$ are binary, and the covariates $\mathbf{S}_{<t} \in \mathcal{S}$ can represent high-dimensional quantities. We formalize this question in the hypothesis

$$
\begin{aligned}
&H_0 : p(\mathbf{s}_{<t}|Y_t = 1) = p(\mathbf{s}_{<t}|Y_t = 0) \quad \text{for all } t \text{ and } \mathbf{s}_{<t} \quad \text{vs.} \\
&H_1 : p(\mathbf{s}_{<t}|Y_t = 1) \neq p(\mathbf{s}_{<t}|Y_t = 0) \quad \text{for some } t \text{ and } \mathbf{s}_{<t}.
\end{aligned} \tag{1}
$$

We assume that the sequence $\{(\mathbf{S}_{<t}, Y_t)\}_{t\geq 0}$ is stationary (Assumption 1) which allows us to rewrite the hypothesis as

$$
\begin{aligned}
&H_0 : p(\mathbf{s}|Y = 1) = p(\mathbf{s}|Y = 0) \quad \text{for all } \mathbf{s} \in \mathcal{S} \quad \text{vs.} \\
&H_1 : p(\mathbf{s}|Y = 1) \neq p(\mathbf{s}|Y = 0) \quad \text{for some } \mathbf{s} \in \mathcal{S}.
\end{aligned} \tag{2}
$$

We will study three different settings (see Figure 2) which admit increasingly complex dependence in $\{\mathbf{S}_{<t}\}_{t\geq 0}$ and $\{Y_t\}_{t\geq 0}$. In Setting A (Figure 2a) there is no temporal dependence, meaning the data $\{(\mathbf{S}_{<t}, Y_t)\}$ are IID. Testing equation (1) is still challenging in this setting when $\mathbf{S}_{<t}$ is high dimensional, but various methods and associated theory have been developed to handle these challenges; see, for example, Kim, Lee and Lei (2019), Kim et al. (2021). In Setting B (Figure 2b) there is temporal dependence in $\{\mathbf{S}_{<t}\}$, but $\{Y_t\}$ are conditionally independent, given the associated value in $\{\mathbf{S}_{<t}\}$. In Setting C (Figure 2c) there




(a) Setting A: $\{(\mathbf{S}_{<t}, Y_t)\}_{t\geq 0}$ with no temporal dependence between pairs $(\mathbf{S}_{<t}, Y_t)$ for different $t$.

(b) Setting B: $Y_t$ conditionally independent of $Y_{t-1}$ given $\mathbf{S}_{<t}$; $\mathbf{S}_{<t}$ is autocorrelated.

(c) Setting C: $Y_t$ conditionally dependent on $Y_{t-1}$ given $\mathbf{S}_{<t}$; $\mathbf{S}_{<t}$ and $Y_t$ are each autocorrelated.

FIG. 2. *Dependence settings. Directed acyclic graphs (DAGs) illustrating the three dependence structures we explore. Note that each variable* $\mathbf{S}_{<t}$ *can itself represent a temporal sequence of high-dimensional functions or images, as in Figure 3.*

is temporal dependence in both $\{\mathbf{S}_{<t}\}$ and $\{Y_t\}$, regardless of the association between the two variables. We expect the TC data to exhibit the structure of Setting C because intensity change labels ($Y_t$) are not conditionally independent solely given the convective structure of the storm ($\mathbf{S}_{<t}$). The effect of convective activity on TC intensity change generally manifests within 24 hours, so a 24-hour history in $\mathbf{S}_{<t}$ should be sufficient (Rogers (2010)).

**3. Sequence data from tropical cyclone satellite imagery.** Analysis of TC convective structure relies on two types of observations: sequences of longwave infrared imagery captured by GOES imagers and records of TC intensity and location recorded in NOAA's HURDAT2 database (Landsea and Franklin (2013)).

Longwave infrared (IR) imagery ($\sim$10.3 $\mu$m wavelength) serves as a proxy for convective strength: where IR-estimated cloud-top temperatures are low, convection is strong. GOES longwave IR imagery is available through NOAA's MERGIR database (Janowiak, Joyce and Xie (2020)) at 30-minute $\times$ 4-km resolution over both the North Atlantic (NAL) and Eastern North Pacific (ENP) basins from 2000–present. Every 30 minutes during the lifetime of a storm, we download a $\sim$2000 km $\times$ 2000 km "stamp" of IR imagery surrounding the TC location. Figure 3 (left) shows two such stamps after an 800-km radius mask is applied.

TC location and intensity are given by the NHC's HURDAT2 best track database which utilizes all available data on each TC (including data not available in real time) to estimate critical characteristics over the lifetime of each TC. HURDAT2 best tracks are provided at six-hour (or synoptic) time resolution; we linearly interpolate TC latitude and longitude between HURDAT2 data points to estimate TC location at nonsynoptic times. Since we are interested in the behavior of mature TCs (as opposed to, e.g., early development), we consider TC genesis to be the first synoptic time at which intensity surpasses 35 kt and lysis to be the last synoptic time at which intensity is at least 35 kt.

*Structural trajectories via ORB.* We have leveraged the ORB framework to analyze the evolution of TC convective structure and demonstrated how projections of ORB functions onto a PCA basis can be used to identify rapid intensification events (McNeely et al. (2020)), but we have not yet directly utilized temporally evolving, continuous functions.

This work studies the temporal evolution of an entire ORB function; in this case the radial profile, defined as $\overline{T}(r) = \frac{1}{2\pi}\int_0^{2\pi} T_b(r,\theta)\,d\theta$. The radial profile $\overline{T}(r)$ captures the structure of cloud-top temperatures $T_b$ as a function of radius $r$ from the TC center and serves as an easily interpretable description of the depth and location of convection near the TC core (McNeely et al. (2020), Sanabia, Barrett and Fine (2014)). The radial profiles are computed at five-km

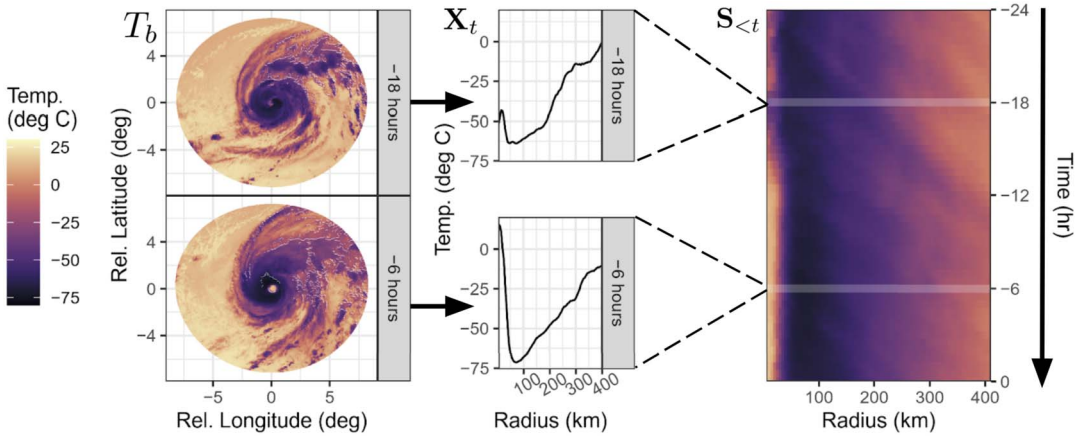



FIG. 3. *Evolution of TC convection as structural trajectories. The raw data for each trajectory* $\mathbf{S}_{<t}$ *is a sequence of a sequence of TC-centered cloud-top temperature images from GOES* ($T_b$). *We convert each GOES image into a radial profile* ($X_t$). *The* 24*-hour sequence of consecutive radial profiles, sampled every* 30 *minutes, defines a* structural trajectory *or Hovmöller diagram* ($\mathbf{S}_{<t}$). *These trajectories serve as high-dimensional inputs to* $\widehat{m}_{\text{post}}(\mathbf{s}_{<t})$.

resolution from zero to 400 km ($d = 80$) (Figure 3, center); we denote these summaries of convective structure at each time $t$ by $\mathbf{X}_t$. Finally, at each time $t$ we stack the preceding 24 hours (48 profiles) into a *structural trajectory* $\mathbf{S}_{<t} = \{\mathbf{X}_t, \mathbf{X}_{t-1}, \ldots, \mathbf{X}_{t-48}\}$. We visualize these trajectories with Hovmöller diagrams (Hovmöller (1949); see Figure 3, right).

*Labeling TC sequence data.* HURDAT2 contains estimated TC intensities only at synoptic times (0000 UTC, 0600 UTC, 1200 UTC and 1800 UTC). We thus begin by labeling these points $Y_t \in \{0, 1\}$, based on whether the TC was undergoing RI (or RW, for those analyses) at time $t$, where $Y = 1$ indicates occurrence of a rapid intensity change event. We then interpolate to nonsynoptic times by assigning label $Y_t = 1$ if the an observation falls between two consecutive synoptic $Y_t = 1$ observations and $Y_t = 0$ otherwise; see McNeely et al. (2023) (Section B) for further details on this procedure.

There are three sample sizes of interest in this application: (i) the number of labeled training sequences $\mathbf{S}_{<t}$ (further divided into 60% train/40% validation), (ii) the number of test sequences $\mathbf{S}_{<t}$ and (iii) the number of synoptic best track entries used when only labels $Y_t$ are required (e.g., $\widehat{m}_{\text{seq}}$ in Algorithm 1(3)). These sample sizes and associated years are summarized in Table 1.

**4. Methods.** Our TC problem setup is difficult because of: (i) the complexity of the data themselves, with one observation representing an entire sequence $\mathbf{S}_{<t}$ of functions, (ii) dependence between labels $Y_t$ (and sequences $\mathbf{S}_{<t}$) at nearby time points $t$ and finally, (iii) the need for scientific interpretability, or more precisely, statistical findings which are easily digestible by TC scientists and forecasters.

We aim to test hypothesis (1) for DID sequence data $\{(\mathbf{S}_{<t}, Y_t)\}_{t\geq 0}$ that satisfy the DAG of Setting C, Figure 2c. Our approach builds on Kim, Lee and Lei (2019), where the authors present a regression approach for detecting differences in high-dimensional IID data $\{(\mathbf{S}_i, Y_i)\}_{i=1}^n$, where $\mathbf{S}_i \in \mathcal{S} := \{\mathbf{s} \in \mathbb{R}^D : p(\mathbf{s}) > 0\}$ for $Y_i \in \{0, 1\}$. Their setup is equivalent to Setting A, Figure 2a. The main idea is to rewrite the two-sample test in the equivalent formulation

$$
\begin{aligned}
&H_0 : \mathbb{P}(Y = 1 | \mathbf{S} = \mathbf{s}) = \mathbb{P}(Y = 1) \quad \text{for all } \mathbf{s} \in \mathcal{S} \quad \text{versus} \\
&H_1 : \mathbb{P}(Y = 1 | \mathbf{S} = \mathbf{s}) \neq \mathbb{P}(Y = 1) \quad \text{for some } \mathbf{s} \in \mathcal{S}.
\end{aligned} \tag{3}
$$

**Algorithm 1:** Test for distributional differences in labeled sequence data

**Require:** type of test (`BST`=TRUE for bootstrap test; `BST`=FALSE for permutation test); train data $\{(\mathbf{S}_{<t}, Y_t)\}_{t\in\mathcal{T}_1}$ and regression method for estimating $m_{\text{post}}(\mathbf{s}) := \mathbb{P}(Y_t = 1|\mathbf{S}_{<t} = \mathbf{s})$; for bootstrap test: train data $\{Y_t\}_{t\in\mathcal{T}_2}$ and regression method for estimating $m_{\text{seq}}(Y_{t-1}, \ldots, Y_{t-k}) := \mathbb{P}(Y_t = 1|Y_{t-1}, \ldots, Y_{t-k})$; number of repetitions $B$; evaluation points $\mathcal{V}$.

**Ensure:** p-value for testing $H_0 : p(\mathbf{s}|Y = 1) = p(\mathbf{s}|Y = 0)$ for all $s \in \mathcal{S}$; local posterior differences $\{\lambda(\mathbf{s})\}_{\mathbf{s}\in\mathcal{V}}$ at the evaluation points $\mathbf{s} \in \mathcal{S}$.

`// Estimate underlying probability distributions`

(1) Estimate $m_{\text{prior}} := \mathbb{P}(Y_t = 1)$ with class proportion $\widehat{m}_{\text{prior}} = \frac{1}{|\mathcal{T}_1|}\sum_{t\in\mathcal{T}_1} Y_t$.

(2) Regress $Y_t$ on $\mathbf{S}_{<t}$ using $\mathcal{T}_1$ to compute $\widehat{m}_{\text{post}}$.

(3) **if** `BST` **then**
Regress $Y_t$ on $Y_{t-1}, \ldots, Y_{t-k}$ using $\mathcal{T}_2$ to compute $\widehat{m}_{\text{seq}}$.
**end if**

`// Compute test statistic and estimate its null distribution`

(4) Compute test statistic $\lambda = \sum_{\mathbf{s}\in\mathcal{V}} \lambda^2(\mathbf{s})$, where $\lambda(\mathbf{s}) = \widehat{m}_{\text{post}}(\mathbf{s}) - \widehat{m}_{\text{prior}}$.

(5) **for** $b \in \{1, 2, \ldots, B\}$ **do**

- Draw new train labels $\{\widetilde{Y}_t\}_{t\in\mathcal{T}_1}$ under $H_0$:
  **if** `BST` **then**
  Draw an initial label sequence $\widetilde{Y}_1, \ldots, \widetilde{Y}_k$ from the empirical distribution.
  Draw sequence of length $100\times k$ from $\widetilde{Y}_t \sim \text{Binom}(\widehat{m}_{\text{seq}}(\widetilde{Y}_{t-k}, \ldots, \widetilde{Y}_{t-1}))$ for burn-in.
  Draw new labels $\widetilde{Y}_t \sim \text{Binom}(\widehat{m}_{\text{seq}}(\widetilde{Y}_{t-k}, \ldots, \widetilde{Y}_{t-1}))$ for $t \in \mathcal{T}_1$.
  **else**
  Permute original labels $\{Y_t\}_{t\in\mathcal{T}_1}$.
  **end if**
- Regress $\widetilde{Y}_t$ on $\mathbf{S}_{<t}$ using $\mathcal{T}_1$ to compute $\widehat{m}^{(b)}_{\text{post}}$.
- Recompute test statistic $\widetilde{\lambda}^{(b)} = \sum_{\mathbf{s}\in\mathcal{V}} \left(\widetilde{\lambda}^{(b)}(\mathbf{s})\right)^2$, where $\widetilde{\lambda}^{(b)}(\mathbf{s}) = \widehat{m}^{(b)}_{\text{post}}(\mathbf{s}) - \widehat{m}_{\text{prior}}$.

`// Compute approximate p-value`

(6) Compute $p$-value according to

$$\widehat{p} = \frac{1}{B+1}\left(1 + \sum_{b=1}^{B} \mathbb{I}\left(\widetilde{\lambda}^{(b)} > \lambda\right)\right).$$

**return** $\widehat{p}$, $\{\lambda(\mathbf{s})\}_{\mathbf{s}\in\mathcal{V}}$

These hypotheses involve a regression function for the "class posterior" $m_{\text{post}}(\mathbf{s}) := \mathbb{P}(Y = 1|\mathbf{S} = \mathbf{s})$ and a "class prior" $m_{\text{prior}} := \mathbb{P}(Y = 1)$. One then tests $H_0$ against $H_1$, using the test statistic

$$\lambda = \sum_{\mathbf{s}\in\mathcal{V}} \left(\widehat{m}_{\text{post}}(\mathbf{s}) - \widehat{m}_{\text{prior}}\right)^2, \tag{4}$$

where $\widehat{m}_{\text{post}}(\mathbf{s})$ is an estimate of $m_{\text{post}}(\mathbf{s})$, $\widehat{m}_{\text{prior}} = \frac{1}{n}\sum_{i=1}^{n} I(Y_i = 1)$ is the class proportion of the training sample and $\mathcal{V} \subset \mathcal{S}$ is a fixed finite set of evaluation data points. Depending on the choice of regression method, the regression test based on $\lambda$ can adapt to challenging nonstandard data, like images and sequences of images or functions.

Because the null distribution of $\lambda$ is typically unknown, Kim, Lee and Lei (2019) compute p-values, based on $\lambda$, by using a permutation procedure. The procedure relies on the exchangeability of the labels $Y$ under $H_0$ in equation (3). However, there are, at least, two types

TABLE 1
*Sample sizes: Data set summary for each category: (i) labeled sequences* $(\mathbf{S}_{<t}, Y_t)$ *used in training, (ii) unlabeled test sequences* $\mathbf{S}_{<t}$ *and (iii) synoptic labels* $Y_t$ *used when complete trajectories are not needed*

| | | NAL | ENP | Total | Year Range | Years |
|---|---|---|---|---|---|---|
| (i) | **Training Data** | | | | | |
| | All Sequences | 47,502 | 31,549 | 79,051 | | |
| | RI Sequences | 7015 | 6742 | 13,757 | | |
| | RW Sequences | 5878 | 7298 | 13,176 | | |
| | **Unique TCs** | 209 | 185 | 394 | 2000–2012 | **13** |
| (ii) | **Test Data** | | | | | |
| | All Sequences | 28,368 | 32,817 | 61,185 | | |
| | RI Sequences | 3965 | 6386 | 10,351 | | |
| | RW Sequences | 3167 | 7182 | 10,349 | | |
| | **Unique TCs** | 125 | 152 | 277 | 2013–2020 | **8** |
| (iii) | **Synoptic Labels** | | | | | |
| | All Labels | 14,683 | 15,274 | 29,957 | | |
| | RI Labels | 1850 | 2462 | 4312 | | |
| | RW Labels | 1643 | 2534 | 4177 | | |
| | **Unique TCs** | 532 | 589 | 1121 | 1979–2012 | **34** |

of dependence in our data $\{(\mathbf{S}_{<t}, Y_t)\}_{t\geq 0}$ which violate the assumption of exchangeability: (i) autocorrelation in $\{Y_t\}_{t\geq 0}$ which is inherent or governed by unobserved quantities, as in Setting C, Figure 2c, and (ii) the presence of an observed, correlated confounding sequence $\{\mathbf{Z}_t\}_{t\geq 0}$ (discussed in Section 8.2). In either case the theoretical guarantees of a valid test in Kim, Lee and Lei (2019) no longer hold.

4.1. *Accounting for dependence in* $Y_t|\mathbf{S}_{<t}$. How do we handle dependence in the labeled sequence data $\{(\mathbf{S}_{<t}, Y_t)\}_{t\geq 0}$? Permutation tests essentially model the distribution of $\{Y_t\}_{t\geq 0}$ assuming IID labels. One way to admit dependence in the relabeling procedure is to instead assume a Markov property of order $k$ on $\{Y_t\}_{t\geq 0}$, that is, to assume that the random variable $Y_t$ depends only on the previous $k$ variables $Y_{t-1}, \ldots, Y_{t-k}$. To estimate the null distribution of the test statistic $\lambda$ (3), we draw new labels from a Markov autoregressive model,

$$\widetilde{Y}_t \sim \text{Bernoulli}\big(\widehat{\mathbb{P}}(Y_t = 1|Y_{t-1} = \widetilde{Y}_{t-1}, \ldots, Y_{t-k} = \widetilde{Y}_{t-k})\big). \tag{5}$$

The marginal distribution of the labels, denoted by $m_{\text{seq}}$, can be estimated from a holdout sample of observed data $\{Y_t\}_{t\geq 0}$ with a variety of methods, including binary Markov chains and random forests. As we shall see, as long as the marginal estimate of $\{Y_t\}_{t\geq 0}$ converges in distribution to the true data-generating process as the size of the holdout sample increases, then the bootstrap test detailed in Algorithm 1 will be asymptotically valid. The result holds, even if $m_{\text{post}}(\mathbf{s})$ and $m_{\text{prior}}$ are not well estimated. This is good news for TC analysis, as one usually has ample access to label series data $Y_t$, whereas sample sizes for the sequences $\mathbf{S}_{<t}$ derived from high-resolution satellite images are smaller.

Proof of the validity of our bootstrap test for equation (3) is given in Section 5, Theorem 1. Section 6 includes empirical results on the power of the test for synthetic data with the DAG structures in Figure 2.

4.2. *Local diagnostics*. Suppose $H_0$ is rejected. That is, we detect that the two distributions $p(\mathbf{s}|Y = 0)$ and $p(\mathbf{s}|Y = 1)$ are indeed different. How do we then provide the scientist with interpretable diagnostics that explain how the two distributions are different?

Classification accuracy tests (Kim et al. (2021)) require a separate post hoc procedure to identify local distributional differences (Chakravarti et al. (2021), Gretton et al. (2012)). A key advantage of the regression test is that the test statistic in equation (4), by construction, is a sum of local posterior differences. Indeed, for each evaluation point $\mathbf{s} \in \mathcal{V}$, we compute the *local posterior difference* (LPD)

$$\lambda(\mathbf{s}) = \widehat{m}_{\text{post}}(\mathbf{s}) - \widehat{m}_{\text{prior}}. \tag{6}$$

A large value of $\lambda(\mathbf{s})$ indicates that the distributions $p(\mathbf{s}|Y=0)$ and $p(\mathbf{s}|Y=1)$ are very different at $\mathbf{s} \in \mathcal{V}$ which, in turn, contributes to a larger test statistic $\lambda = \sum_{\mathbf{s}\in\mathcal{V}} \lambda^2(\mathbf{s})$ and potential rejection of $H_0$.

REMARK 1. The posterior difference $\lambda(\mathbf{s}) = \mathbb{P}(Y=1|\mathbf{s}) - \mathbb{P}(Y=1)$ can be viewed as a *scaled density difference*,

$$\lambda(\mathbf{s}) = \frac{p(\mathbf{s}|Y=1) - p(\mathbf{s}|Y=0)}{w(\mathbf{s})}, \tag{7}$$

where $w(\mathbf{s}) = \frac{1}{1-\pi} p(\mathbf{s}|Y=1) + \frac{1}{\pi} p(\mathbf{s}|Y=0)$ is a positive scaling function and $\pi := \mathbb{P}(Y=1)$ denotes the prior class probability or $m_{\text{prior}}$. This difference has several desired properties for assessing local distributional differences: $\lambda(\mathbf{s})$ is always bounded, unlike other popular discrepancy measures, such as the density ratio, $p(\mathbf{s}|Y=1)/p(\mathbf{s}|Y=0)$, and the density difference, $p(\mathbf{s}|Y=1) - p(\mathbf{s}|Y=0)$, itself. Furthermore, the posterior difference does not decay to zero as fast as $p(\mathbf{s})$ which leads to high sensitivity to detect differences in low density regions; for example, in the case of balanced classes, $\lambda(\mathbf{s})$ takes a value of $+\frac{1}{2}$, when $p(\mathbf{s}|Y=1) \gg p(\mathbf{s}|Y=0)$, and a value of $-\frac{1}{2}$, when $p(\mathbf{s}|Y=1) \ll p(\mathbf{s}|Y=0)$, regardless of the actual magnitudes of $p(\mathbf{s})$, $p(\mathbf{s}|Y=1)$, and $p(\mathbf{s}|Y=0)$.

In summary, our method tests for distributional differences in labeled sequence data as follows:

1. Decide on a suitable model for the marginal distribution of $\{Y_t\}_{t\geq 0}$.
2. Apply Algorithm 1 to compute the p-value for testing the hypotheses in equation (3).
3. If $H_0$ is rejected, then examine local posterior differences (6) to identify what patterns $\mathbf{s}$ in the state space $\mathcal{S}$ of sequence data contributed the most to the rejection.

**5. Theory.** This section provides theoretical justification for our bootstrap procedure for testing equation (3). In particular, we show that Algorithm 1 controls the type I error asymptotically.

ASSUMPTION 1 (Stationary sequence). $\{(\mathbf{S}_{<t}, Y_t)\}_{t\geq 0}$ is a stationary sequence, where $\mathbf{S}_{<t} \in \mathcal{S}$ and $Y_t \in \{0, 1\}$

Assumption 1 is needed for the hypothesis in equation (2) to be well defined.

ASSUMPTION 2 (Conditional independence). $\{(\mathbf{S}_{<t}, Y_t)\}_{t\geq 0}$ satisfies the DAG of Setting C (Figure 2).

Assumption 2 encodes the conditional independences required for our method to control type I error.

In this section we denote the test statistic by

$$\lambda(\mathcal{D}) = \int \left(\widehat{m}_{\text{post}}(\mathbf{s}) - \widehat{m}_{\text{prior}}\right)^2 dQ(\mathbf{s}), \tag{8}$$

where $Q$ is any fixed measure over $\mathcal{S}$, and $\widehat{m}_{\text{post}}$ and $\widehat{m}_{\text{prior}}$ are obtained using a training set $\mathcal{D} := \{(\mathbf{S}_{<t}, Y_t)\}_{t\in\mathcal{T}_1}$. In Algorithm 1, $Q$ is the distribution that assigns mass $1/|\mathcal{V}|$ for each evaluation point $\mathbf{s} \in \mathcal{V}$, but the results we show here apply to any $Q$. We also assume that the regression estimator we use is a continuous function of the training set.

ASSUMPTION 3 (Continuous regression method). $\widehat{m}_{\text{post}}$ is obtained by applying a regression estimator that is a continuous function of $\mathcal{D}$.

Moreover, let $\mathcal{D}_0^{t_2} := \{(\mathbf{S}_{<t}, Y_t^0)\}_{t\in\mathcal{T}_1}$ denote a random draw from the data set used to estimate the null distribution of $\lambda$, where $\{Y_t^0\}_{t\in\mathcal{T}_1} \sim \mathrm{G}_{\widehat{\mathbf{p}}_{t_2}}$ and G is a distribution over $\{0,1\}^{|\mathcal{T}_1|}$, indexed by the parameter $\widehat{\mathbf{p}}_{t_2}$, which is estimated using a holdout set $\mathcal{D}' = \{Y_t\}_{t\in\mathcal{T}_2}$ with $t_2 = |\mathcal{T}_2|$. In the method described in Section 4.1, $\mathrm{G}_{\widehat{\mathbf{p}}_{t_2}}$ is the Markov autoregressive model $\widehat{m}_{\text{seq}}$, but this model can be more general. We require it to converge to the true distribution of the marginal process $\{Y_t\}_{t\in\mathcal{T}_1}$ when the null hypothesis holds:

ASSUMPTION 4 (Consistency of the marginal distribution estimator). The estimator $\widehat{\mathbf{p}}_{t_2}$ is such that if the null hypothesis is true,

$$\mathrm{G}_{\widehat{\mathbf{p}}_{t_2}} \xrightarrow[t_2\longrightarrow\infty]{\text{Dist}} \mathrm{G}^*,$$

where $\mathrm{G}^*$ is the true generating process of $\{Y_t\}_{t\in\mathcal{T}_1}$.

In the following we show two examples where Assumption 4 holds.

EXAMPLE 1. Under Settings A and B (Figure 2), $Y_t$'s are IID under the null hypothesis. Thus, $G^*$ is necessarily a product of IID Bernoulli random variables with some parameter $p$. Now, let $G_{\widehat{\mathbf{p}}_{t_2}}$ be the product of IID Bernoulli random variables with parameter given by $p_{t_2} := (t_2)^{-1}\sum_{t\in\mathcal{T}_2} Y_t$. The law of large numbers implies that $p_{t_2} \xrightarrow[t_2\longrightarrow\infty]{\text{a.s.}} p$. Thus, the cumulative distribution function of $Y_t^0$, given by

$$F_{Y_t^0}(y_t) = \begin{cases} 0 & \text{if } y_t < 0, \\ 1 - p_{t_2} & \text{if } 0 \le y_t < 1, \\ 1 & \text{otherwise} \end{cases}$$

is such that $F_{Y_t^0}(y_t) \xrightarrow[t_2\longrightarrow\infty]{} F_{Y_t}(y_t)$. It follows that

$$\begin{aligned}\mathbb{P}(Y_t^0 \le y_t, \forall t \in \mathcal{T}_1) &= \prod_{t\in\mathcal{T}_1} F_{Y_t^0}(y_t) \\ &\xrightarrow[t_2\longrightarrow\infty]{} \prod_{t\in\mathcal{T}_1} F_{Y_t}(y_t) = \mathbb{P}(Y_t \le y_t, \forall t \in \mathcal{T}_1),\end{aligned}$$

and, therefore, Assumption 4 holds.

EXAMPLE 2 (Markov Chain). If (under $H_0$) the process $\{Y_t\}_{t\ge 0}$ is an irreducible and ergodic stationary $k$-order Markov chain, then the maximum likelihood estimators of the transition probabilities converge almost surely to the true transition probabilities (Grimmett and Stirzaker (2020)). The same reasoning of Example 1 then implies that Assumption 4 holds for such estimator under Setting C.

The following theorem shows that, under $H_0$, the test statistic has approximately the same distribution as the test statistic evaluated at the generated data $\mathcal{D}_0^{t_2}$.

THEOREM 1. *Assume 1, 2, 3 and 4. Under the null hypothesis,*

$$\lambda\big(\mathcal{D}_0^{t_2}\big) \xrightarrow[t_2 \longrightarrow \infty]{\text{Dist}} \lambda(\mathcal{D}).$$

It follows from Theorem 1 that type I error is controlled asymptotically.

COROLLARY 1 (Type I error control). *Let*

$$\widehat{p}_B^{t_2}(\mathcal{D}) := \frac{1}{B+1}\left(1 + \sum_{b=1}^{B} \mathbb{I}\big(\lambda\big(\mathcal{D}^{(b)}\big) > \lambda(\mathcal{D})\big)\right)$$

*be the Monte Carlo p-value for $H_0$, where $\mathcal{D}^{(1)}, \ldots, \mathcal{D}^{(B)} \overset{IID}{\sim} \mathcal{D}_0^{t_2}$. Assume that Assumptions 1, 2, 3 and 4 hold. Then, under the null hypothesis, for any $0 < \alpha < 1$,*

$$\lim_{t_2 \longrightarrow \infty} \lim_{B \longrightarrow \infty} \mathbb{P}\big(\widehat{p}_B^{t_2}(\mathcal{D}) \le \alpha\big) = \alpha.$$

See McNeely et al. (2023) (Section C) for proofs.

**6. Performance of tests on synthetic data.** In this section we use synthetic data to examine the performance (validity, power and diagnostics) of our Markov chain bootstrap test for the data dependence settings in Figure 2c.

6.1. *Synthetic univariate sequence data.* For simplicity, we first consider a scalar covariate $S_t$ of interest. We then create dependent sequences $\{S_t, Y_t\}_{t \ge 0}$ with a logistic generative model,

$$\begin{aligned} Y_t | S_t &\sim \text{Bernoulli}(p_t), \\ p_t &= \text{logistic}\big(\gamma H_\delta(S_t) + U_t\big), \\ H_\delta(S) &= \begin{cases} 0 & |S| < \delta, \\ S & |S| \ge \delta, \end{cases} \\ S_t &= U_t', \qquad U_t' \sim \text{AR}_{\phi'}(1), \\ U_t &\sim \text{AR}_{\phi}(1). \end{aligned} \tag{9}$$

The variables $U_t$ and $U_t'$ are spurious variables (not included in the DAGs), which induce autocorrelation in the binary response variable $Y_t$ and the covariate $S_t$, respectively. In our toy example, we assume that both $U_t$ and $U_t'$ are given by autoregressive models of order 1; more specifically, by AR(1) models of the form $U_t = \phi U_{t-1} + \sqrt{1-\phi^2}\epsilon_t$ and $U_t' = \phi' U_{t-1}' + \sqrt{1-\phi'^2}\epsilon_t'$, where $\epsilon_t, \epsilon_t' \overset{\text{iid}}{\sim} N(0,1)$ and $\phi, \phi' \in [0,1]$ are parameters for the 1-lag autocorrelation in $U_t$ and $U_t'$, respectively. Increasing $\phi'$ thus increases the autocorrelation (but not the variance) of the variable of interest $S_t$, while increasing $\phi'$ increases the autocorrelation (but not the variance) of the spurious variable $U_t$.

The parameter $\gamma \ge 0$ determines the signal strength, or the strength of the dependence of $Y_t$ on $S_t$. Testing $H_0$ in equation (1) is equivalent to testing $H_0 : \gamma = 0$. Ideally, our method should also be able to identify local regions in the sample space $\mathcal{S}$ where the two distributions are different or the same. To assess our method's local performance, we hence include a hard thresholding operator $H_\delta(\cdot)$ in equation (9), which, regardless of the signal strength $\gamma$, creates a region in $\mathcal{S} = \mathbb{R}$, where $\mathbb{P}(Y = 1 | S = s) = \mathbb{P}(Y = 1)$ for $s \in (-\delta, \delta)$.

The two parameters $\phi$ and $\phi'$ allow us to create synthetic data with the dependence structures in Figure 2. More specifically:

*Setting A*: $\phi' = \phi = 0$.

*Setting B*: $\phi' > 0$ induces autocorrelation in $\{S_t\}_{t\geq 0}$ via $U'_t$; $\phi = 0$.

*Setting C*: $\phi' > 0$ induces autocorrelation in $\{S_t\}_{t\geq 0}$ via $U'_t$, while $\phi > 0$ induces autocorrelation in $\{Y_t\}_{t\geq 0}$ via $U_t$.

REMARK 2. In this toy example, $S_t = U'_t$ with $U'_t$ observed. More generally, $\mathbf{S}_{<t}$ can depend on unmeasured variables $\mathbf{U}'_t$ as well as confounding variables $\mathbf{Z}_t$ which are connected to both $\mathbf{S}_{<t}$ and $Y_t$. The latter setting is discussed in Section 8.2. We also note that the spurious variables $\mathbf{U}_t$ and $\mathbf{U}'_t$ can have more complex temporal dependence (than in the example), as indicated by the fully connected sequences in Figure 2c.

6.2. *Test results for synthetic data*. The logistic generative model in equation (9) provides a variety of controls over the dependence structure of $\{(S_t, Y_t)\}_{t\geq 0}$. For our synthetic experiments we implement Algorithm 1 with either the MC bootstrap or the permutation test for $|\mathcal{V}| = 250$ evaluation points. We estimate the regression function $m_{\text{post}}(s) = \mathbb{P}(Y_t = 1|S_t = s)$ using train data $\{(S_t, Y_t)\}_{t\in\mathcal{T}_1}$ and a Nadaraya–Watson (NW) kernel estimator with an Epanechnikov kernel and the bandwidth chosen as the sample standard deviation of $s$ divided by $|\mathcal{T}_1|^{1/5}$ (Li and Racine (2007)). For the bootstrap test we estimate the label distribution $m_{\text{seq}}$ using an order $k = 4$ Markov chain and labels $\{Y_t\}_{t\in\mathcal{T}_2}$.

*Validity*: Testing $H_0$ in equation (1) is equivalent to testing $H_0 : \gamma = 0$ (no signal strength). To examine validity, we set $\gamma = 0$ and simulate 500 independent data sets for each experiment, or combination of "setting" (A, B, C) and "test method" (permutation or MC bootstrap test with $|\mathcal{T}_1| = |\mathcal{T}_2| = |\mathcal{V}| = 250$). That is, each experiment returns 500 p-values. If the test controls type I error, we expect these p-values to be approximately uniformly distributed. Figure 4 assesses validity by plotting the difference between the empirical and (uniform) theoretical quantiles against the theoretical quantiles; this is equivalent to a standard quantile-

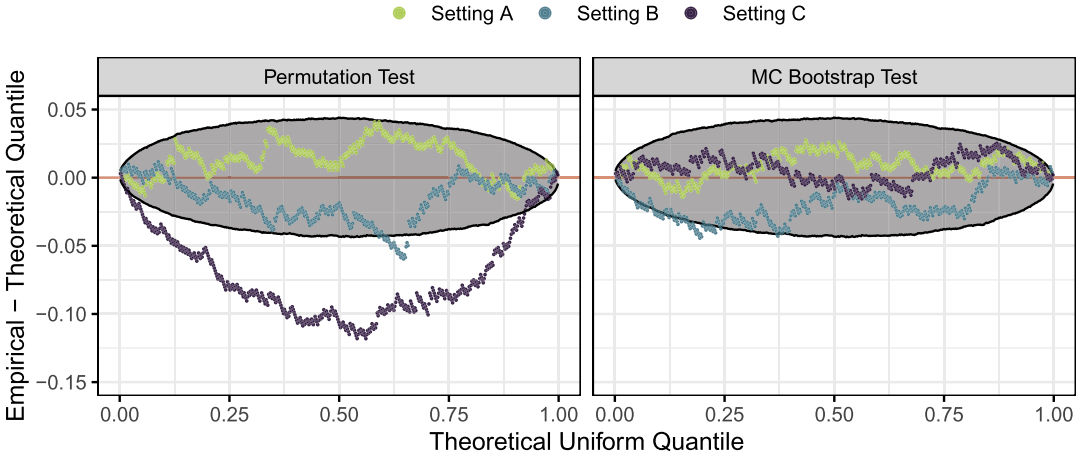



FIG. 4. *Synthetic sequence data*: *Validity of permutation test* (*left*) *vs. Markov chain bootstrap test* (*right*). *Under* $H_0 : \gamma = 0$, *a valid test is expected to return uniformly distributed p-values. Each curve corresponds to a different experimental setting* (*A, B or C*), *and shows the difference between empirical and uniform theoretical quantiles for* 500 *repetitions*; *see text for details. The gray region represents a* 95% *pointwise confidence interval derived from Monte Carlo samples of* 500 *uniform deviates. Under Setting C* (*labels dependent even after conditioning on predictors*; *purple curve*), *the permutation test* (*left panel*) *does not control the type I error, but the Markov chain bootstrap test* (*right panel*) *does*. (*Setting A*: $\phi = \phi' = 0$. *Setting B*: $\phi = 0$, $\phi' = 0.8$. *Setting C*: $\phi = \phi' = 0.8$. *Sample sizes* $|\mathcal{T}_1| = |\mathcal{T}_2| = |\mathcal{V}| = 250$.)

quantile plot with the diagonal subtracted. As a baseline, we provide a 95% confidence interval of this difference based on 10,000 Monte Carlo simulations of 500 uniformly distributed random variables.

The permutation test (left panel) is valid under Settings A and B, where $Y_t$ and $Y_{t-1}$ are independent after conditioning on $S_t$. However, under Setting C, the p-values tend to have lower values than a uniform distribution, corresponding to higher-than-nominal type I errors at most significance levels. The MC bootstrap test (right panel) controls the type I error at all significance levels for all three dependence settings, indicating that our adjustment to account for the dependence in $Y_t \mid S_t$ achieved the desired result.

*Power*: We next examine how the power of the test $H_0 : \gamma = 0$ vs. $H_1 : \gamma \neq 0$ depends on:

i. signal strength $\gamma$ (Figure 5, top),
ii. train sample size $|\mathcal{T}_1|$ (Figure 5, bottom),
iii. autocorrelation $\phi$ in labels $\{Y_t\}_{t\geq 0}$ or, equivalently, correlation in $\{Y_t|S_t\}_{t\geq 0}$ (Figure 6, left) and
iv. autocorrelation $\phi'$ in predictors $\{S_t\}_{t\geq 0}$ (Figure 6, right).

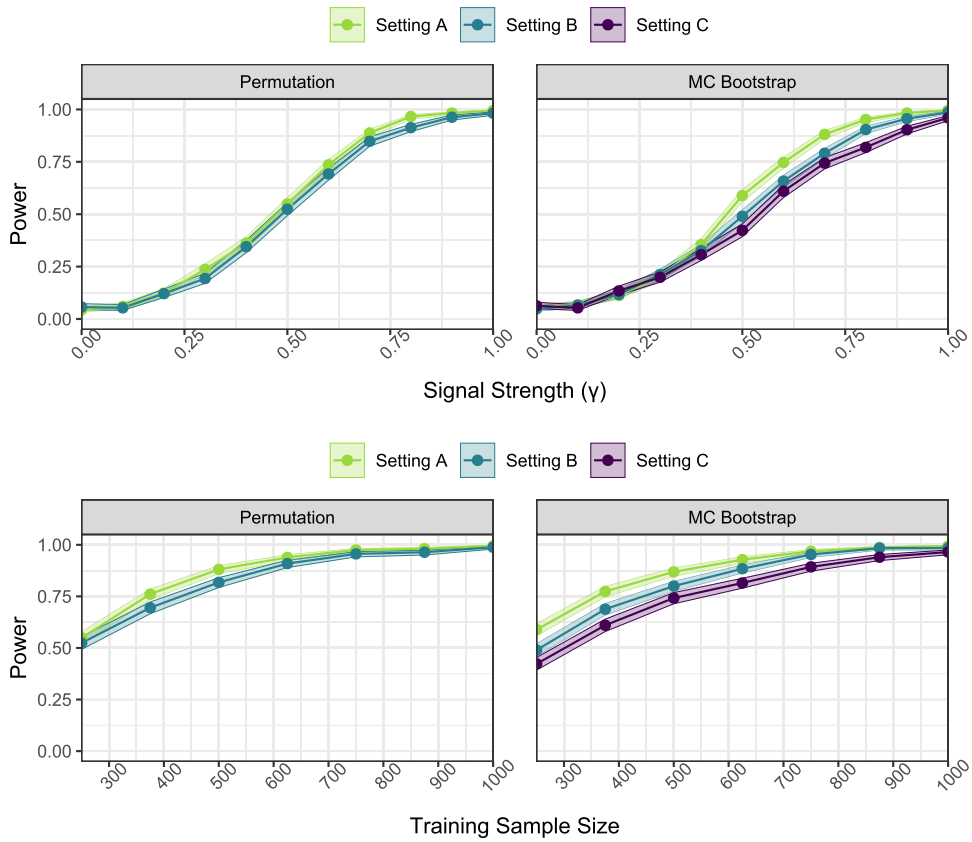



FIG. 5. *Synthetic sequence data*: *Power as a function of signal strength and training sample size*. Top: *The power of all tests increases with the signal strength* $\gamma$, *regardless of dependence setting. The MC bootstrap test has similar power as the permutation test, but the former test can be applied to the more challenging Setting C with dependent labels* $Y_t|S_t$. *Sample sizes* $|\mathcal{T}_1| = |\mathcal{T}_2| = |\mathcal{V}| = 250$. Bottom: *The power of all tests, at the alternative* $\gamma = 0.5$, *increases with the train sample size* $|\mathcal{T}_1|$, *regardless of dependence setting. Sample sizes* $|\mathcal{T}_2| = |\mathcal{V}| = 250$. *The filled regions represent* 95% *pointwise confidence intervals for binomial proportions.* (*Setting A*: $\phi = \phi' = 0$. *Setting B*: $\phi = 0, \phi' = 0.8$. *Setting C*: $\phi = \phi' = 0.8$.) *Setting C is not shown for the permutation test, as it is not valid.*

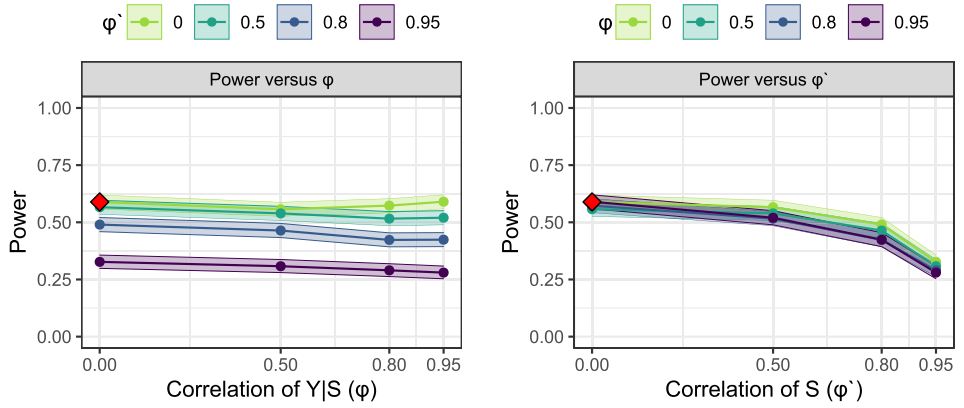



FIG. 6. *Synthetic sequence data*: *Power of bootstrap test as function of autocorrelation of* $Y$ (left) *and* $S$ (right). *The red diamond-shaped markers correspond to Setting A with* `IID` *data. The power to detect the signal* $\gamma = 0.5$ *is independent of the correlation in* $Y_t|S_t$ (*value of* $\phi$) *but decreases with correlation in* $S_t$ (*larger values of* $\phi'$). *The filled regions represent* 95% *pointwise confidence intervals for binomial proportions.* (*Sample sizes* $|\mathcal{T}_1| = |\mathcal{T}_2| = |\mathcal{V}| = 250$.)

At each fixed value of $\gamma$, we perform 1000 simulations. Power is then estimated as the fraction of rejected null hypotheses at the $\alpha = 0.05$ level. To ensure validity, we choose $|\mathcal{T}_1| \geq 250$, as before for the MC bootstrap test. (The permutation test is valid by construction under Settings A and B, but not Setting C.)

As expected, the power increases with the signal strength $\gamma$ for all tests and dependence settings (Figure 5, top). When both tests are valid, the MC test has the same power as the permutation test. The practical implication is that, even if one *thinks* Setting B is a good approximation to the problem at hand, there are benefits to applying the MC bootstrap test: one can achieve similar power with the advantage of having robustness in the event that the labels are dependent after conditioning on predictors.

Figure 5 (bottom) indicates that the power of the tests may be determined by the *quality of the regression estimator* $\widehat{m}_{\text{post}}(\mathbf{s})$: indeed, Figure 5 (bottom) shows that the power at a fixed alternative ($\gamma = 0.5$) increases with the train sample size $|\mathcal{T}_1|$. The latter result is consistent with Theorem 3.3 of Kim, Lee and Lei (2019), which states for a regression permutation test under Setting A, that if the chosen regression method $\widehat{m}_{\text{post}}(\mathbf{s})$ has a small mean integrated squared error, then the power of testing (2) is large over a wide region of alternative hypotheses.

Figure 6 brings insight on how dependence in $\{(S_t, Y_t)\}_{t \geq 0}$ affect the power of the MC bootstrap test. The red diamond-shaped markers represent Setting A with `IID` sequence data ($\phi = \phi' = 0$). Increasing correlation in the labels $Y_t|S_t$ (larger values of $\phi$) has no effect on power, while the test remains valid as long as $\widehat{m}_{\text{seq}}$ is accurate; this further emphasizes the previous result that the bootstrap test is robust to correlation in $Y_t|S_t$ without sacrificing power. Meanwhile, increasing correlation in $S_t$ (larger values of $\phi'$) reduces power; this follows from a reduced effective sample size which, in turn, reduces the quality of $\widehat{m}_{\text{post}}$.

*Local posterior differences*: For our synthetic example the hard thresholding operator $H_\delta(\cdot)$ induces a region $s \in (-\delta, +\delta)$, where $p(s|Y = 1) = p(s|Y = 0)$. If the null hypothesis in equation (2) is rejected, then the estimated LPDs can identify the regions of large vs. small distributional differences, as long as the regression estimator $\widehat{m}_{\text{post}}$ is consistent and the train sample size $|T_1|$ is sufficiently large. Figure 7 shows the average and one standard deviation estimates of the LPD for a NW kernel estimator over 200 simulations.

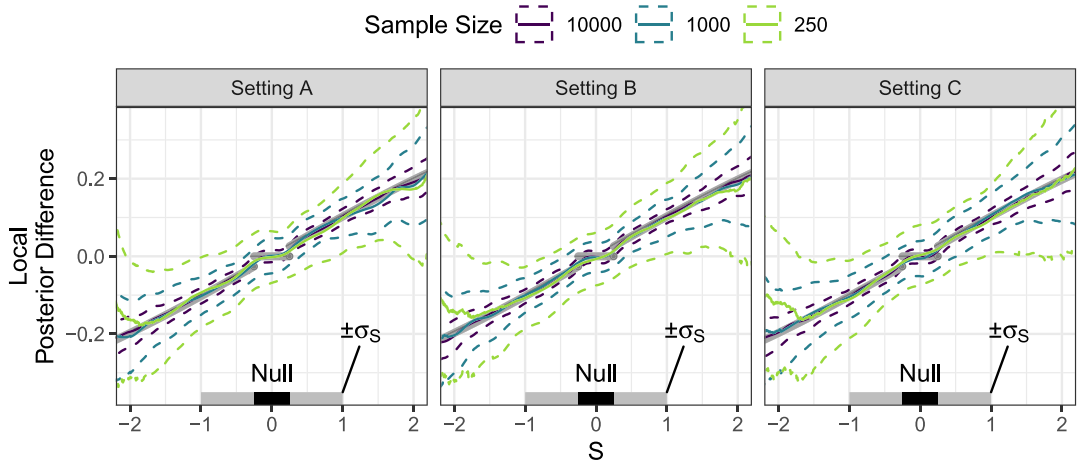



FIG. 7. *Synthetic sequence data*: *Local posterior difference in null region. The gray curves in each panel/setting show the true LPD which is zero in the local null (no signal) region* $s \in (-0.25, 0.25)$. *The solid and dashed colored curves represent the mean and one standard deviation estimates of the LPD over* 200 *simulated data sets. These estimates are, on average, close to the true LPD, with the dispersion decreasing for a local nonparametric estimator (like the NW kernel estimator) as the train sample size increases. The gray bar at the bottom marks the standard deviation of s; the large variance of the estimated LPDs far from* $s = 0$ *is partially due to the concentration of data near* $s = 0$. (*Setting A*: $\phi = \phi' = 0$. *Setting B*: $\phi = 0$, $\phi' = 0.8$. *Setting C*: $\phi = \phi' = 0.8$. *Sample sizes* $|\mathcal{T}_1|$ *vary*, $|\mathcal{T}_2| = |\mathcal{V}| = 250$.)

**7. Relating evolution of TC convection to rapid intensity change.** In our TC study, each observation consists of: (i) a 24-hour sequence $\mathbf{S}_{<t} = \{\mathbf{X}_t, \mathbf{X}_{t-1}, \ldots, \mathbf{X}_{t-48}\}$ of one-dimensional radial profile functions $\mathbf{X}_t = \frac{1}{2\pi}\int_0^{2\pi} T_{b,t}(r, \theta)\, d\theta$, sampled every 30 minutes for a total of 48 profiles (see Figure 3) and (ii) a binary label $Y_t \in \{0, 1\}$ for the entire sequence.[2] Since all individual sample points that are part of a rapid intensity change event are labeled as $Y = 1$, a 24-hour sequence $\mathbf{S}_{<t}$ with sequence label $Y_t = 1$ could either be part of an ongoing RI/RW event (if $Y_t = 1$ falls near the end of an event) or be part of the lead-up to RI/RW (if $Y_t = 1$ falls near the beginning of an event). Analyses of the latter case, such as approaches that can identify archetypal modes of structural evolution preceding the *onset* of RI/RW, are particularly valuable to forecasting of RI/RW events.

We divide our TC study into three parts:

1. Analysis by event type (RI vs. not RI, or RW vs. not RW) within each basin, North Atlantic (NAL) or eastern North Pacific (ENP) for a total of four different two-sample tests.
2. Case studies of three tropical cyclones (Hurricanes Nicole, Jose and Rosa).
3. Analysis of a subset of our data that consists of sequences immediately preceding RI onset. (We ask whether our regression two-sample test can find archetypical evolutionary modes preceding RI onset, and if so, what the "lead time" between typical patterns and the RI onset would be.)

As in the synthetic example, we estimate $m_{\text{seq}}$ in Algorithm 1 with a Markov chain of order $k = 8$. The sequence data $\mathbf{S}_{<t}$ are, however, much more complex than in our synthetic example. This is where we benefit from a more complex regression method for estimating $m_{\text{post}}$;

[2]We check that the dominant principal components of $\mathbf{X}_t$ and the continuous intensities used to derive $Y_t$ are stationary via augmented Dickey–Fuller tests; the p-value of each test (including tests of the first three ORB coefficients in McNeely et al. (2020)) are all $< 10^{-20}$. We conclude that Assumption 1 is reasonable for these data.

here, we fit a convolutional neural network to the 24-hour sequence data. Further details on how we estimate $m_{\text{seq}}$ and $m_{\text{post}}$ can be found in McNeely et al. (2023) (Section A).

7.1. *Analysis by event type and basin. Significance test.* We start by testing $H_0$ : $p(\mathbf{s}_t|Y_t = 1) = p(\mathbf{s}_t|Y_t = 0)$ by event type and basin. For rapid intensification (RI) the MC bootstrap test rejects $H_0$ at level 0.05 for both the NAL and ENP basins, meaning that we indeed detect a significant difference ($p < 0.01$) in 24-hour sequences $\mathbf{S}_{<t}$ of convective structure leading up to RI vs. not-RI events. For rapid weakening (RW) the MC bootstrap test rejects $H_0$ in the ENP basin, but not in the NAL basin.

Our results are consistent with scientists' understanding of TCs; for rapid intensification, a TC exhibits a narrow range of convective patterns "primed" to efficiently convert heat energy to mechanical energy across the storm, hence the structural difference in convection for RI vs. not-RI events. Rapid weakening, on the other hand, is a more complex process driven by several factors external to the TC, such as vertical wind shear, which may not be fully captured by convective structure. In addition, RW is expected to be more difficult to detect in the NAL basin due to the broader range of possible environmental configurations and the increased rarity of over-water RW in the basin.

*Local posterior difference.* Next, we investigate what kind of structural patterns lead to the rejection of $H_0$ for the RI-NAL, RI-ENP and RW-ENP models. Figure 8 (left) shows a two-dimensional embedding of the sequence data via principal component analysis (PCA: computed separately for each basin). Each point represents a 24-hour sequence $\mathbf{S}_{<t}$ colored by its local posterior different (LPD). Note that PCA is only used for purposes of visualization; the test itself is performed on the entire sequence of radial profiles without a prior dimension reduction step. Figure 8 (right) shows examples of Hovmöller diagrams for six 24-hour sequences $\mathbf{S}_{<t}$ sorted by LPD and TC intensity.

TCs are known to have different distributions of $\mathbf{S}_{<t}$ for different basins. Nevertheless, we identify the same type of evolutionary patterns of convective structure for RI-NAL (panel i) and RI-ENP (panel ii): Positive LPD or "high chance of RI" (see diagrams A–C for i and ii) tends to occur for cold cloud tops near the core (dark blue at smaller radius), growing in coverage and depth of convection with time (dark blue region extending to larger radii when going from −24 to 0 hours). Meanwhile, negative LPD or "low chance of RI" (see diagrams D–F for i and ii) tends to occur when TCs already possess a well-defined eye (narrow yellow region near the center) or exhibit decaying core convection (dark blue region decreasing in size when going from −24 to 0 hours). While such patterns can be directly quantified and studied in future works, this work remains focused on exploration of entire radial profiles.

Finally, RW-ENP results (panel iii) are not exactly opposite of the RI-ENP results (panel ii), meaning that "high chance of RW" patterns might not mirror "low chance of RI" patterns, and vice versa. In particular, TCs commonly form strong convective cores without eyes (iii-E) *prior* to intensification, form an eye (end of i-C) during RI, then rapidly weaken by dissipating entirely (ii-D, iii-A, iii-B) without reforming a cold, eyeless core because the reduction of intensity is accompanied by a collapse of convection throughout the TC. Thus, RI/RW-ENP results are not symmetric. Unfortunately, the RW-ENP model also predominantly captures the trivial result that currently-intense TCs are more likely to weaken; see Section 8.2 for a discussion of potential corrections.

7.2. *Hurricane case studies.* Thus far, we have analyzed the collection of 24-hour sequences in the 2013–2020 test sample as a whole; however, forecasters monitor *individual* storms in real time for signals of RI. Here, we take an in-depth look at our results for three individual TCs in the test set: Hurricanes Nicole, Jose and Rosa. Each of these storms display distinct evolutionary modes. We track each TC through its lifetime to investigate the relationship between the evolution of convective structure, the LPDs and intensity change.

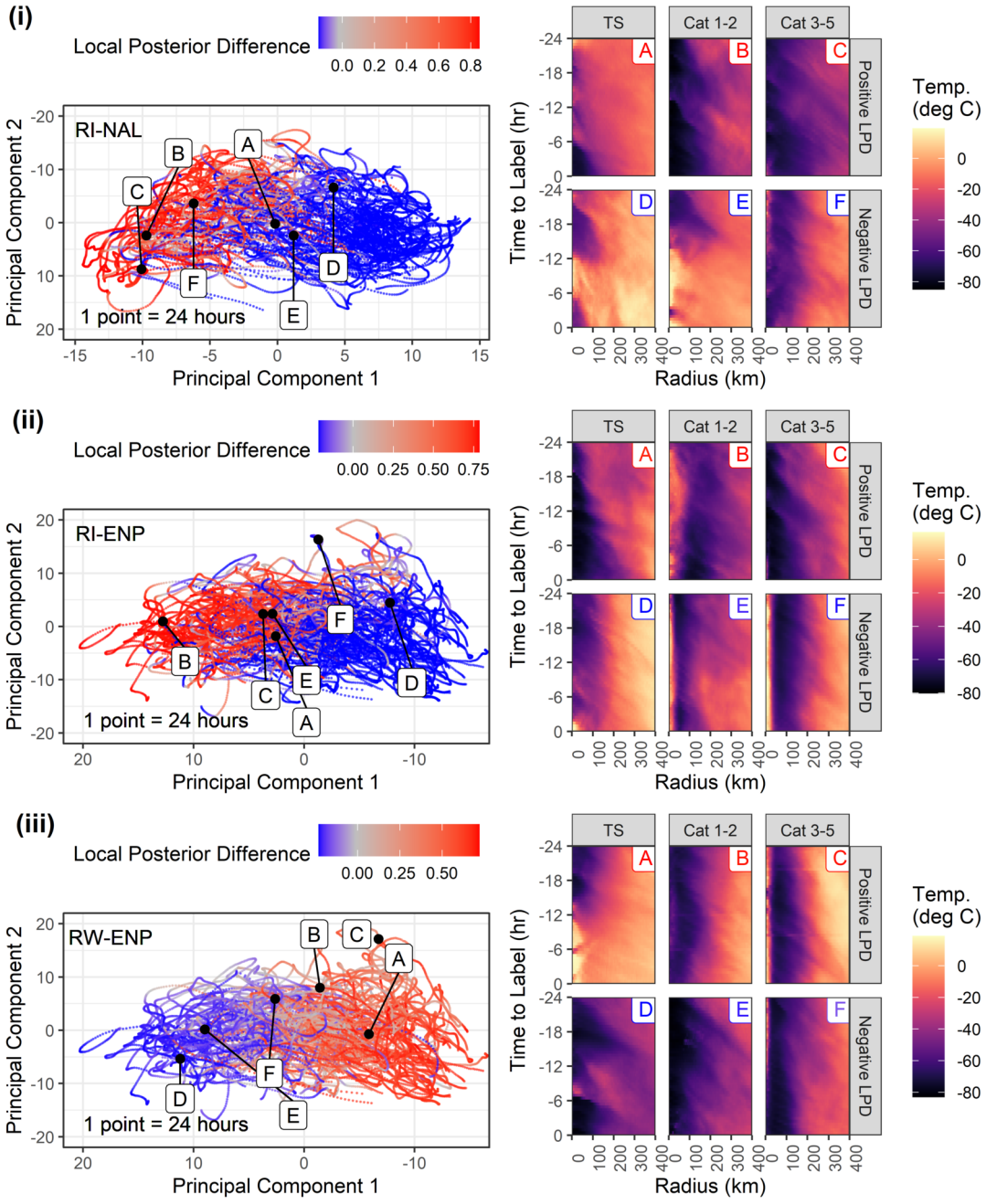



FIG. 8. *Analysis by event type and basin. The MC bootstrap test rejects* $H_0$ *(i.e., it detects significant differences in convection) for the RI-NAL, RI-ENP and RW-ENP models; the test is not rejected for the RW-NAL model. Analyses of local posterior difference (LPD) are shown for the first three models (see panels i–iii).* Left column: *Two-dimensional PCA map of sequence data. One point in the map represents a* 24*-hour structural trajectory (sequence of radial profiles)* $\mathbf{S}_{<t}$ *with the color coding for the estimated LPD.* Right column: *Six* 24*-hour structural trajectories at locations* A–F *in the PCA map, shown as Hovmöller diagrams (recall Figure 3). The examples for each study are selected at random from each combination of LPD sign (positive LPD:* A–C*; negative LPD:* D–F*) and TC intensity (Tropical Storm:* A*,* D*; Category* 1–2*:* B*,* E*; Category* 3–5*:* C*,* F*); see text for a discussion of the results.*

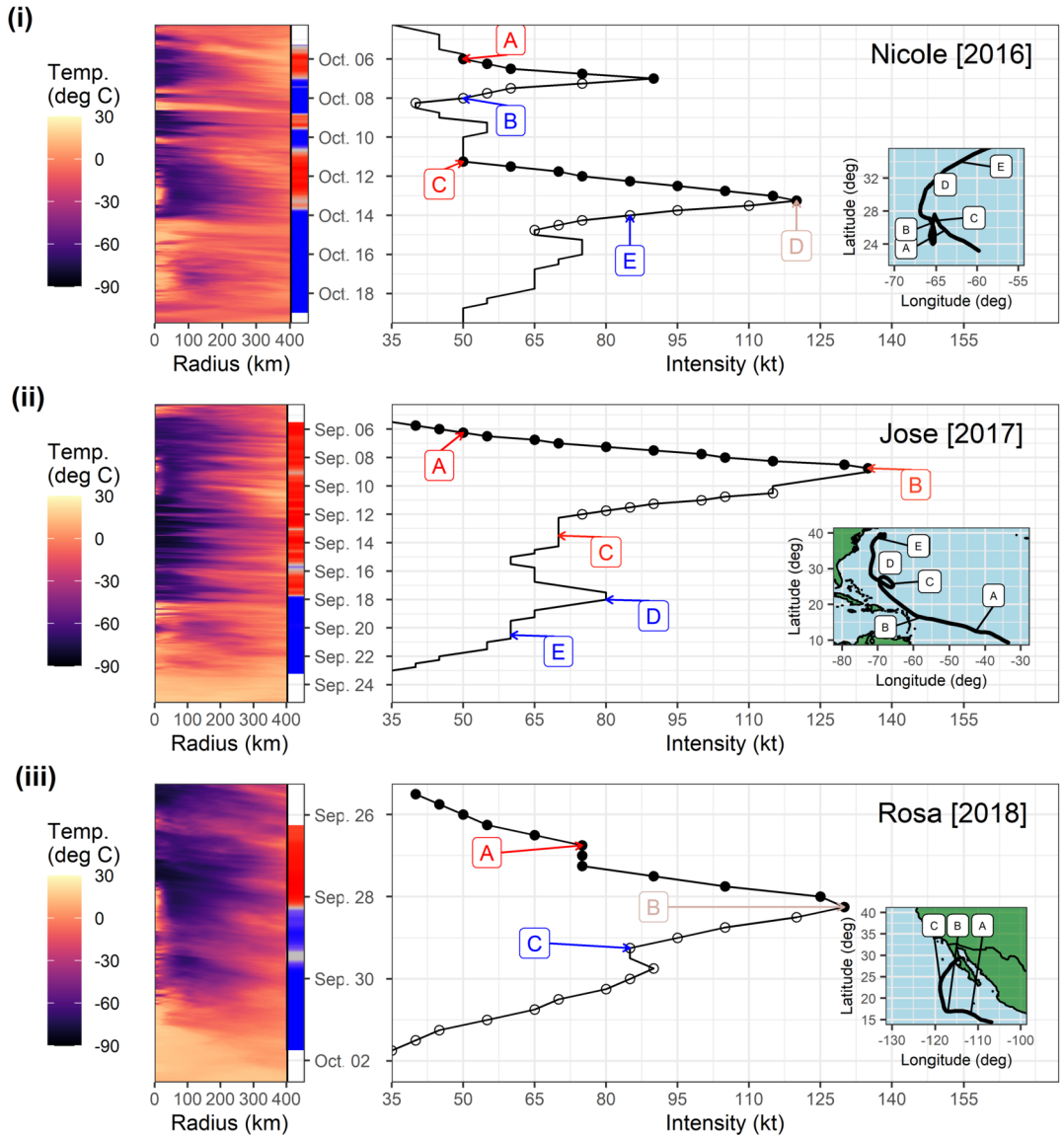



FIG. 9. *Case studies of Hurricanes Nicole* [*NAL* 2016, *i*], *Jose* [*NAL* 2017, *ii*] *and Rosa* [*ENP* 2018, *iii*]. Left: *Structural trajectories of each storm through its entire lifetime, with the sidebar to the right showing the LPD at each* 30*-minute observation; each LPD value is evaluated using the preceding* 24*-hour sequence for an RI model* (*trained on the train sample for that basin*) *with positive values in red and negative values in blue*. Right: *Storm intensity over time, where filled vs. empty circles mark RI and RW events, respectively. The physical track of the storm across the North Atlantic is shown inset, with the same time points labeled*. Panel (i): *For Hurricane Nicole, the LPDs capture rapidly intensifying periods* (A, C, D), *collapsing convection* B *and the decay of the TC eye* E. Panel (ii): *Hurricane Jose was subjected to high vertical wind shear near September* 9, *which our model does not account for; while the core convection of the TC remained poised for RI, high shear instead caused rapid weakening*. Panel (iii): *Hurricane Rosa exhibited two interesting phenomena captured by the LPDs. First, it experienced a pause in its rapid intensification* A; *the LPDs indicate an ongoing RI threat at this time, consistent with the resumption of intensification* 18 *hours later. Second, beginning on September* 28, *the TC underwent an eyewall replacement cycle, associated with weakening. The LPDs mark this shift at the TC's peak intensity as well as the brief period* (*following* C) *where eyewall replacement is completed prior to landfall.*

Figure 9 (left) shows the evolution of radial profiles for each storm. Appended to the right of these profiles is a sidebar showing the LPDs from the RI model for the associated basin. The right panels display the evolution of each storm's intensity, where filled and hollow markers indicate RI and RW events, respectively. The storm's physical track over the ocean is shown as an inset. These three hurricane case studies respectively highlight: (i) signals which lead RI, (ii) the effect of vertical wind shear and (iii) the appearance of an eyewall replacement cycle (a process by which a second eyewall forms, robbing the TC of energy as it shrinks to replace the original eyewall).

Panel (i) depicts *Hurricane Nicole* [2016]. The TC underwent RI [A] on October 6 before its intensity stalled while influenced by the outflow from Hurricane Matthew [2016] which induced vertical wind shear [B]. Several days later, Nicole reintensified [C], [D] before again weakening and then transitioning into an extratropical system [E]. The LPDs in Figure 9 appear to align with intensity changes and the emergence of deep convection [A], [C] and eye formation [D]. The structural trajectories immediately preceding [A] and [C] are particularly interesting: the TC has not yet begun to intensify rapidly, but 24-hour sequences, including and prior to [A] and [C], have strongly positive LPDs. These results indicate that structural trajectories of radial profiles in the North Atlantic may contain signals of RI prior to onset.

Panel (ii) shows *Hurricane Jose* [2017] and highlights the importance of vertical wind shear. This TC exhibited deep convection in the core for nearly two weeks, remaining at elevated RI risk according to the RI-NAL posterior differences. However, after an initial period of RI, high vertical wind shear disrupted the TC structure around September 9. The TC decayed from 135 kt to about 70 kt and never appreciably intensified again, despite several periods of elevated LPDs prior to [D]. The underlying regression $\widehat{m}_{\text{post}}$ does not account for the vertical wind shear which prevented the TC's intensification; see Section 8.2 for a discussion about how to potentially account for external factors such as vertical wind shear.

Finally, panel (iii) depicts the short-lived eastern North Pacific TC *Hurricane Rosa* [2018] which underwent an extended period of RI before beginning an eyewall replacement cycle. This evolution included two interesting phenomena. First, the TC experienced a pause in its rapid intensification [A]; the LPDs indicate an ongoing RI threat at this time, consistent with the resumption of intensification 12–18 hours later. Second, beginning on September 28, the TC underwent an eyewall replacement cycle which manifests as a expansion of the eye accompanied by an evening out of convection across the storm. Such cycles typically result in a decrease of the TC's maximum sustained winds. The LPDs mark this shift at the TC's peak intensity as well as the brief period [C] where eyewall replacement is completed prior to landfall.

7.3. *Structural trajectories preceding rapid intensification.* In the previous section, case studies indicated that positive LPDs can *lead* RI events. That is, they may signal convective structure primed for RI *prior* to RI onset. That core convection can predict RI is known; at least one RII forecast model, used by the NHC, includes the fraction of GOES pixels between 50 and 200 km with temperatures below $-30°$C as a predictor (Kaplan et al. (2015)). We would thus expect our LPDs to contain some signal leading RI.

In this section we hone in on a subset of the RI-NAL test sample, which represents complete 24-hour sequences $\mathbf{S}_{<t}$, leading up to RI onset points (such as points i-[A] and i-[C] in Figure 9). This leaves us with a total of 36 sequences, visualized in the PCA map of Figure 10 (left). We then investigate whether we are able to detect RI within the 48 hours prior to onset, and if so, what the lead time might be. As before, a positive posterior difference indicates $p(\mathbf{s}_t|Y_t = 1) > p(\mathbf{s}_t|Y_t = 0)$. Because $\mathbf{S}_{<t}$ here encodes the recent evolution of convective

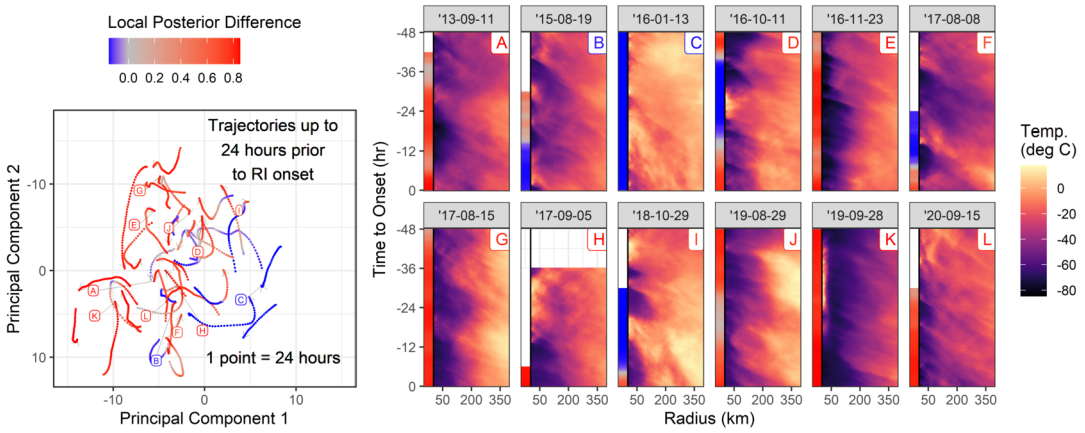



FIG. 10. *Detecting RI prior to onset.* Left: *PCA map of a subset of the RI-NAL test sample, representing* 24*-hour sequences prior to RI onset. We show* 36 *events with at least* 24 *hours of nonmissing radial profiles* $X_t$ (*i.e., at least one sequence* $\mathbf{S}_{<t}$). Right: *Examples of* 48*-hour structural trajectories preceding RI onset. The local posterior differences, shown in the bar to the left of each map, largely signal TCs with deep core convection in the last* 24 *hours as "primed for RI."*

structure, we interpret LPD $\lambda(\mathbf{s}_t) \gg 0$ prior to RI as an indication that the TC at time $t$ exhibits convective structures primed for RI onset. That is, $\lambda(\mathbf{s}_t) \gg 0$ in these cases indicates an *elevated risk of RI* based on convective evolution alone.

In 30 (83%) out of 36 onsets, the RI-NAL LPDs indicate an above-average RI threat at onset. In 28 (78%) and 24 (66%) onsets, LPDs indicate an above-average RI threat six or 12 hours prior to onset, respectively. Some example trajectories are shown in Figure 10 (right). Prior to RI onset, TCs tend to exhibit deep core convection, evidence of strong outflow (indicated by the downward slant of $\mathbf{S}_{<t}$) and often pronounced diurnal cycles (oscillations in $\mathbf{S}_{<t}$ over time). These results indicate that structural trajectories or LPDs from our test could serve as valuable inputs to RI forecast models such as the SHIPS-RII (Kaplan et al. (2015)).

## 8. Limitations and potential extensions of methods.

8.1. *Adding multiple data sources, environmental variables and other functional features.* Deep convection, as revealed by IR imagery, is only one ingredient required for RI onset. Our current analysis does not include environmental variables, such as vertical wind shear, which are known to affect RI. Our analysis also neglects asymmetric patterns in the radial structure of convection and other key ORB functional features. Although these features were omitted from the presented analysis, our testing framework for distributional differences of a binary response variable can be extended to handle these settings with multiple data sources and different data types.

For a joint analysis of several variables, additional predictors can be added to the regression model $\widehat{m}_{\text{post}}$ so that the sequence $\mathbf{S}_{<t}$ consists of, for example, both structural features (e.g., ORB functions) and environmental features (e.g., vertical wind shear or sea surface temperatures). Furthermore, the CNN model can easily handle multiple inputs, including other observation bands (e.g., the water vapor-sensitive 6.9 $\mu$m GOES band or microwave imagery from polar-orbiting satellites) and ORB functions, other than radial profiles (e.g., the size of temperature level sets as a function of the temperature threshold; see McNeely et al. (2019), McNeely et al. (2020)). These inputs could be included as isolated channels that combine at either the dense layer (with separate feature extraction layers) or at the logistic layer (with separate feature extraction and dense layers).

In particular, the current TC analysis of GOES IR-imagery can be extended to capture asymmetry in convective structure by separating the radial profiles into four quadrants. Our ongoing work on structural forecasting illustrates the promise of such an approach (McNeely et al. (2022)).

8.2. *Adjusting for other variables*. Our presented framework detects potential associations between high-dimensional time series and binary labels. Any detected association could be caused by variables that confound the relationship between the label and the covariates of interest. Our methodology can be adjusted to account for confounding variables by: (1) generalizing the test of independence to a test of conditional independence and (2) including additional covariates in the regression problem:

Suppose that we want to detect distributional differences in sequence data $\{\mathbf{S}_{<t}\}_{t\geq 0}$, preceding an event $Y_t = 1$ vs. a nonevent $Y_t = 0$, after adjusting for the effect of other variables with sequence data $\{\mathbf{Z}_{<t}\}_{t\geq 0}$. For example, wind shear might confound the relationship between convective structures and RI or RW. Assuming a stationary process $\{(\mathbf{S}_{<t}, \mathbf{Z}_{<t}, Y_t)\}_{t\geq 0}$, hence omitting $t$, we test conditional independence

$$\begin{aligned} &H_0 : p(\mathbf{s}|Y = 1, \mathbf{Z} = \mathbf{z}) = p(\mathbf{s}|Y = 0, \mathbf{Z} = \mathbf{z}) \quad \text{for all } \mathbf{s} \in \mathcal{S} \text{ and all } \mathbf{z} \quad \text{vs.} \\ &H_1 : p(\mathbf{s}|Y = 1, \mathbf{Z} = \mathbf{z}) \neq p(\mathbf{s}|Y = 0, \mathbf{Z} = \mathbf{z}) \quad \text{for some } \mathbf{s} \in \mathcal{S} \text{ or } \mathbf{z}. \end{aligned} \tag{10}$$

These hypotheses are equivalent to

$$\begin{aligned} &H_0 : \mathbb{P}(Y = 1|\mathbf{S} = \mathbf{s}, \mathbf{Z} = \mathbf{z}) = \mathbb{P}(Y = 1|\mathbf{Z} = \mathbf{z}) \quad \text{for all } \mathbf{s} \in \mathcal{S} \text{ and all } \mathbf{z} \quad \text{vs.} \\ &H_1 : \mathbb{P}(Y = 1|\mathbf{S} = \mathbf{s}, \mathbf{Z} = \mathbf{z}) \neq \mathbb{P}(Y = 1|\mathbf{Z} = \mathbf{z}) \quad \text{for some } \mathbf{s} \in \mathcal{S} \text{ or } \mathbf{z}. \end{aligned} \tag{11}$$

Analogous to equation (4), we can define a regression test statistic $\lambda$ based on the difference between an estimate of $\mathbb{P}(Y = 1|\mathbf{S} = \mathbf{s}, \mathbf{Z} = \mathbf{z})$ and $\mathbb{P}(Y = 1|\mathbf{Z} = \mathbf{z})$. Note that if $\mathbf{Z}$ is associated with both $\mathbf{S}$ and $Y$, then a permutation test is not valid, even for Setting A with IID data, because of lack of exchangeability under $H_0$. One solution for (IID as well as DID) sequence data $\{(\mathbf{S}_{<t}, \mathbf{Z}_{<t}, Y_t)\}_{t\geq 0}$ is to extend our bootstrap test to a procedure where one estimates the distribution of the label series $\{Y_t\}_{t\geq 0}$ *conditional on* $\mathbf{Z}$.

Regarding TC rapid intensity change, the admission of confounders would improve the interpretability of both the RI and RW tests. In the case of RI, wind shear ($\mathbf{Z}$) is a powerful environmental predictor and can inhibit intensification $Y$ of a TC with otherwise favorable structure $\mathbf{S}$ (e.g., Hurricane Jose [2017], Figure 9, ii-C). Meanwhile, the results for RW (Figure 8(iii)) appear to weakly capture the obvious relationship: stronger storms are more likely to rapidly weaken. By accounting for the effect of current intensity ($\mathbf{Z}$), we could better assess the relationship between structural evolution ($\mathbf{S}$) and intensity change ($Y$).

8.3. *Local P-values*. In this work we refer to the local posterior difference (equation (6)) as a local diagnostic, rather than as a local p-value, because empirical results show that we do not control the type I error of a pointwise test $H_0(\mathbf{s}) : \mathbb{P}(Y = 1|\mathbf{S} = \mathbf{s}) = \mathbb{P}(Y = 1)$ at current sample sizes. The LPD value can, however, in principle, be used to test the local null

$$H_0^{\epsilon}(\mathbf{s}) : \mathbb{P}\big(Y = 1|\mathbf{S} = \mathbf{s}'\big) = \mathbb{P}(Y = 1) \quad \text{for all } \mathbf{s}' \in B(\mathbf{s}; \epsilon), \tag{12}$$

where $B$ is a ball of radius $\epsilon$ centered at $\mathbf{s}$. In McNeely et al. (2023) we show that, under DAG B, the local p-values (12) are valid if the regression estimator $\widehat{m}_{\text{post}}$ only uses the observations in $\mathcal{D}$ such that $\mathbf{S} \in B(\mathbf{s}; \epsilon)$. The latter assumption holds for regression estimators that are based on partitions, such as tree-based estimators (random forests, boosting methods) as well as smoothing kernel estimators with finite support.

8.4. *Bootstrapping the label series.* To estimate the null distribution of the test statistic $\lambda$ (3), we currently assume a $k$-step Markov chain and draw new labels from the Markov autoregressive model (5). There are other ways one can bootstrap the label distribution, including adopting sampling schemes that model long-range dependence in the label sequences. For a review of bootstrap methods for dependent data, see, for example, Bühlmann (2002), Horowitz (2003), and Kreiss and Paparoditis (2011).

**9. Conclusions.** We describe a statistical framework for analyzing the relationship between complex high-dimensional data $\{\mathbf{X}_t\}_{t\geq 0}$ and labels $\{Y_t\}_{t\geq 0}$. For DID sequence data $\{(\mathbf{S}_t, Y_t)\}_{t\geq 0}$, where $\mathbf{S}_{<t} = \{\mathbf{X}_{t-T}, \mathbf{X}_{t-T+1}, \ldots, \mathbf{X}_t\}$ and $T > 0$, we propose a two-sample test (equation (2) and Algorithm 1) with minimal assumptions beyond stationarity. The test relies on two simple key ideas: (i) a test statistic based on the posterior difference $\mathbb{P}(Y = 1|\mathbf{S}) - \mathbb{P}(\mathbf{S})$, which we estimate using a machine learning algorithm suitable for the data at hand (empirical results indicate that the test power depends on the quality of the regression estimate, Section 6 and Figure 5) and (ii) a bootstrap test, where we estimate the marginal distribution of $\{Y_t\}_{t\geq 0}$ (consistency guarantees asymptotic validity, Theorem 1). Our framework provides interpretable diagnostics in local posterior differences (Section 7) and can be extended to include longer-range dependence structures (Section 8.4), multiple data sources (Section 8.1) and potential confounding variables (Section 8.2).

9.1. *TC results.* We detect a distributional difference between sequences leading up to RI vs. not-RI events in both the North Atlantic and eastern North Pacific basins ($p < 0.01$). Local posterior differences for RI-NAL and RI-ENP indicate that specific types of convection—deep and deepening core convection—are present both before and during RI (Figures 8 and 9). Furthermore, we observe that particular convective structures are necessary for RI (Figure 10) and thus useful indicators of future RI, but they are not sufficient to trigger RI on their own (Figure 9, ii), as the TC environment (e.g., vertical wind shear and ocean heat content) must also support intensification. Thus, while our current results have apparent value for RI forecasting, an analysis of structural trajectories alongside environmental factors, such as vertical wind shear, promises better understanding of RI and may improve analysis of RW events as well.

When posing the same question regarding RW vs. not-RW events, we do not detect a difference in the NAL basin ($p = 0.18$); this is expected, as RW is more likely to be driven by a variety of internal and external factors not well captured by convective distribution, whereas RI generally *requires* TC convective structure to be capable of sustaining rapid energy uptake. However, the ENP basin is characterized by a narrow spatial region of conditions favorable to TC development such as warm ocean waters and moist air; this homogeneity leads to a significant signal for RW in the ENP basin ($p < 0.01$).

9.2. *Broader impact.* While we apply our methods to meteorology, the proposed statistical framework is applicable to any labeled DID sequence data. Labeled video or other sequence data are common in automation, medical monitoring and multiple domains in the physical sciences; in many of these areas, the ability to identify high-risk patterns in spatiotemporal data could prove transformative. The flexibility of our framework in admitting a fusion of multiple data sources and adjustment for other variables is also vital to analyzing the complex systems in the environmental and physical sciences and many other domains.

**Funding.** This work is supported in part by NSF Grant DMS-2053804 and NSF PHY-2020295.

RI is grateful for the financial support of CNPq (309607/2020-5) and FAPESP (2019/11321-9).

## SUPPLEMENTARY MATERIAL

**Supplemental text** (DOI: 10.1214/22-AOAS1668SUPPA; .pdf). The online supplement contains a description of our CNN regressor in Section 7, our algorithm for labeling RI and RW events, and a proof of Theorem 1.

**Code files** (DOI: 10.1214/22-AOAS1668SUPPB; .zip). All code used to produce the results in this paper is available at https://github.com/ihmcneely/ORB2sample and as online supplementary material. All data used are publicly available: HURDAT2 at https://www.nhc.noaa.gov/data/#hurdat and MERGIR at https://disc.gsfc.nasa.gov/datasets/GPM_MERGIR_1/summary, which can both be accessed via provided code.

## REFERENCES


AMINIKHANGHAHI, S. and COOK, D. J. (2017). A survey of methods for time series change point detection. *Knowl. Inf. Syst.* **51** 339–367. https://doi.org/10.1007/s10115-016-0987-z

BERRETT, T. B., WANG, Y., BARBER, R. F. and SAMWORTH, R. J. (2020). The conditional permutation test for independence while controlling for confounders. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **82** 175–197. MR4060981

BÜHLMANN, P. (2002). Bootstraps for time series. *Statist. Sci.* **17** 52–72. MR1910074 https://doi.org/10.1214/ss/1023798998

CANDÈS, E., FAN, Y., JANSON, L. and LV, J. (2018). Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 551–577. MR3798878 https://doi.org/10.1111/rssb.12265

CHAKRAVARTI, P., KUUSELA, M., LEI, J. and WASSERMAN, L. (2021). Model-independent detection of new physics signals using interpretable semi-supervised classifier tests. arXiv preprint arXiv:2102.07679.

DEMARIA, M. and KAPLAN, J. (1999). An updated statistical hurricane intensity prediction scheme (SHIPS) for the Atlantic and eastern North Pacific basins. *Weather Forecast.* **14** 326–337.

EVANS, C. and G'SELL, M. (2020). Sequential changepoint detection for label shift in classification. arXiv preprint arXiv:2009.08592.

GALL, R., FRANKLIN, J., MARKS, F., RAPPAPORT, E. N. and TOEPFER, F. (2013). The hurricane forecast improvement project. *Bull. Am. Meteorol. Soc.* **94** 329–343.

GRETTON, A., BORGWARDT, K. M., RASCH, M. J., SCHÖLKOPF, B. and SMOLA, A. (2012). A kernel two-sample test. *J. Mach. Learn. Res.* **13** 723–773. MR2913716

GRIMMETT, G. R. and STIRZAKER, D. R. (2020). *Probability and Random Processes*. Oxford Univ. Press, Oxford. Fourth edition [of 0667520]. MR4229142

HOROWITZ, J. L. (2003). Bootstrap methods for Markov processes. *Econometrica* **71** 1049–1082. MR1995823 https://doi.org/10.1111/1468-0262.00439

HOVMÖLLER, E. (1949). The trough-and-ridge diagram. *Tellus* **1** 62–66.

JANOWIAK, J., JOYCE, B. and XIE, P. (2020). NCEP/CPC L3 half hourly 4 km global (60S–60N) merged IR V1. https://doi.org/10.5067/P4HZB9N27EKU

KANDASAMY, K., KRISHNAMURTHY, A., POCZOS, B., WASSERMAN, L. A. and ROBINS, J. M. (2015). Nonparametric von Mises estimators for entropies, divergences and mutual informations. In *NIPS* **15** 397–405.

KAPLAN, J. and DEMARIA, M. (2003). Large-scale characteristics of rapidly intensifying tropical cyclones in the North Atlantic basin. *Weather Forecast.* **18** 1093–1108.

KAPLAN, J., DEMARIA, M. and KNAFF, J. A. (2010). A revised tropical cyclone rapid intensification index for the Atlantic and eastern North Pacific basins. *Weather Forecast.* **25** 220–241.

KAPLAN, J., ROZOFF, C. M., DEMARIA, M., SAMPSON, C. R., KOSSIN, J. P., VELDEN, C. S., CIONE, J. J., DUNION, J. P., KNAFF, J. A. et al. (2015). Evaluating environmental impacts on tropical cyclone rapid intensification predictability utilizing statistical models. *Weather Forecast.* **30** 1374–1396.

KATSEVICH, E. and RAMDAS, A. (2020). A theoretical treatment of conditional independence testing under model-x. arXiv preprint arXiv:2005.05506v4.

KIM, I., LEE, A. B. and LEI, J. (2019). Global and local two-sample tests via regression. *Electron. J. Stat.* **13** 5253–5305. MR4043073 https://doi.org/10.1214/19-EJS1648

KIM, I., RAMDAS, A., SINGH, A. and WASSERMAN, L. (2021). Classification accuracy as a proxy for two-sample testing. *Ann. Statist.* **49** 411–434. MR4206684 https://doi.org/10.1214/20-AOS1962

KLOTZBACH, P. J., BOWEN, S. G., PIELKE, R. and BELL, M. (2018). Continental U.S. hurricane landfall frequency and associated damage: Observations and future risks. *Bull. Am. Meteorol. Soc.* **99** 1359–1376. https://doi.org/10.1175/BAMS-D-17-0184.1

KNAFF, J. A. and DEMARIA, R. T. (2017). Forecasting tropical cyclone eye formation and dissipation in infrared imagery. *Weather Forecast*. **32** 2103–2116.

KREISS, J.-P. and PAPARODITIS, E. (2011). Bootstrap methods for dependent data: A review. *J. Korean Statist. Soc*. **40** 357–378. MR2906623 https://doi.org/10.1016/j.jkss.2011.08.009

LANDSEA, C. W. and FRANKLIN, J. L. (2013). Atlantic hurricane database uncertainty and presentation of a new database format. *Mon. Weather Rev.* **141** 3576–3592. https://doi.org/10.1175/MWR-D-12-00254.1

LI, Q. and RACINE, J. S. (2007). *Nonparametric Econometrics*: *Theory and Practice*. Princeton Univ. Press, Princeton, NJ. MR2283034

LUO, C., LOU, J.-G., LIN, Q., FU, Q., DING, R., ZHANG, D. and WANG, Z. (2014). Correlating events with time series for incident diagnosis. In *Proceedings of the* 20*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1583–1592.

MCNEELY, T., LEE, A. B., HAMMERLING, D. and WOOD, K. (2019). Quantifying the spatial structure of tropical cyclone imagery.

MCNEELY, T., LEE, A. B., WOOD, K. M. and HAMMERLING, D. (2020). Unlocking GOES: A statistical framework for quantifying the evolution of convective structure in tropical cyclones. *J. Appl. Meteorol. Climatol*. **59** 1671–1689.

MCNEELY, T., KHOKHLOV, P., DALMASSO, N., WOOD, K. M. and LEE, A. B. (2022). Structural forecasting for short-term tropical cyclone intensity guidance. arXiv preprint arXiv:2206.00067.

MCNEELY, T., VINCENT, G., IZBICKI, R., WOOD, K. M. and LEE, A. B. (2023). Supplement to "Detecting distributional differences in labeled sequence data with application to tropical cyclone satellite imagery." https://doi.org/10.1214/22-AOAS1668SUPPA, https://doi.org/10.1214/22-AOAS1668SUPPB

MOON, K. and HERO, A. (2014). Multivariate f-divergence estimation with confidence. *Adv. Neural Inf. Process. Syst*. **27** 2420–2428.

ROGERS, R. (2010). Convective-scale structure and evolution during a high-resolution simulation of tropical cyclone rapid intensification. *J. Atmos. Sci*. **67** 44–70.

SANABIA, E. R., BARRETT, B. S. and FINE, C. M. (2014). Relationships between tropical cyclone intensity and eyewall structure as determined by radial profiles of inner-core infrared brightness temperature. *Mon. Weather Rev.* **142** 4581–4599. https://doi.org/10.1175/MWR-D-13-00336.1

SCHARWÄCHTER, E. and MÜLLER, E. (2020a). Does terrorism trigger online hate speech? On the association of events and time series. *Ann. Appl. Stat*. **14** 1285–1303. MR4152133 https://doi.org/10.1214/20-AOAS1338

SCHARWÄCHTER, E. and MÜLLER, E. (2020b). Two-sample testing for event impacts in time series. In *Proceedings of the* 2020 *SIAM International Conference on Data Mining* 10–18. SIAM.

SESIA, M., SABATTI, C. and CANDÈS, E. J. (2019). Gene hunting with hidden Markov model knockoffs. *Biometrika* **106** 1–18. MR3912377 https://doi.org/10.1093/biomet/asy033

WOOD, K. M. and RITCHIE, E. A. (2015). A definition for rapid weakening of North Atlantic and eastern North Pacific tropical cyclones. *Geophys. Res. Lett*. **42** 10,091–10,097.