Detecting Distributional Differences in Labeled Sequences of Tropical Cyclone Satellite Imagery

Ann B. Lee
Department of Statistics & Data Science
Carnegie Mellon University
(Joint work with Trey McNeely, Galen Vincent, Kim Wood, and Rafael Izbicki)

12/5/20

2023 AOAS Paper

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

Trey McNeely
(PhD 2022, CMU)

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*

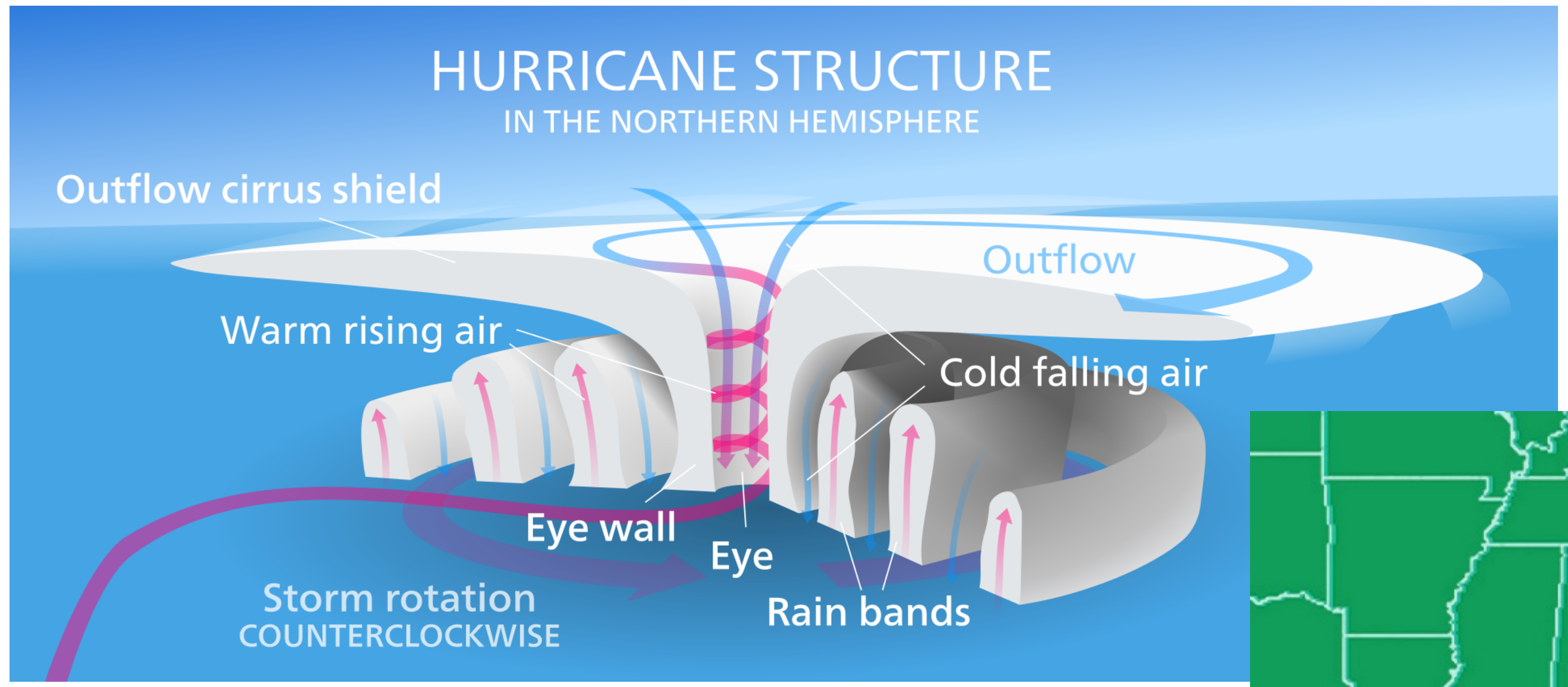
Both Methodology and Applied Relevance

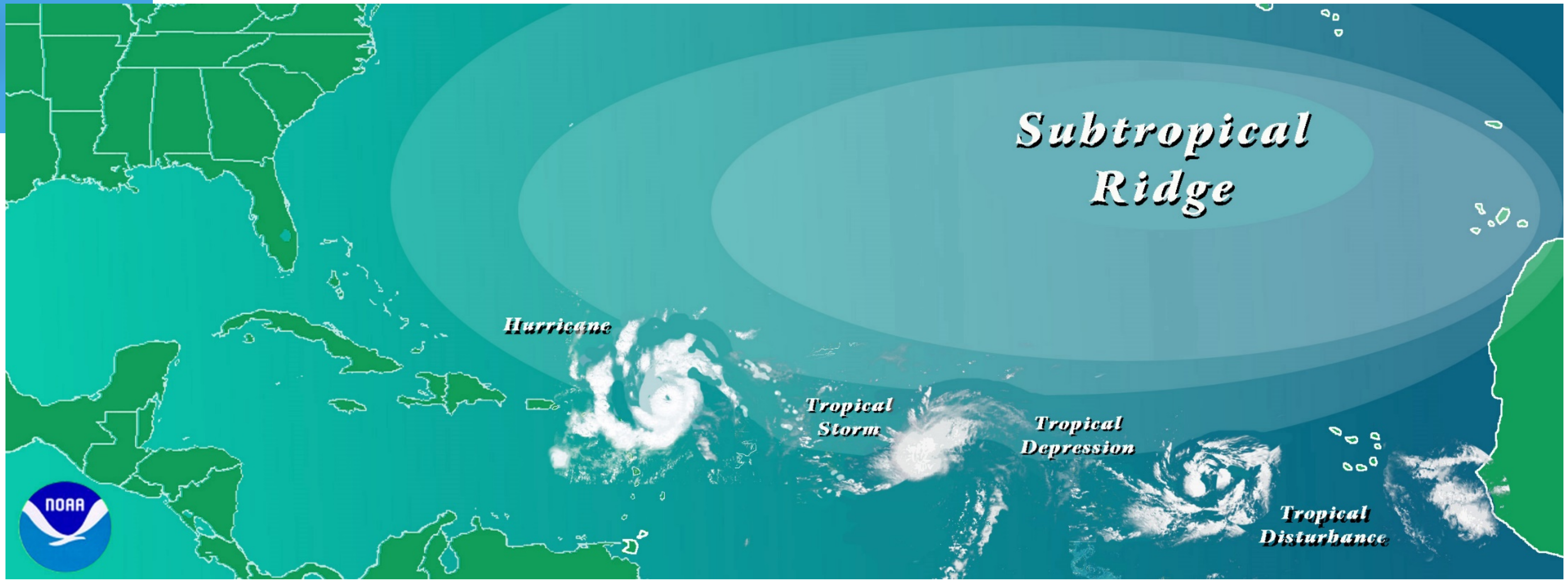
- Motivating application:
  - To identify spatiotemporal patterns in tropical cyclone (TC) satellite imagery that lead up to an upcoming rapid intensity change event.
- Requires new methodology:
  - For detecting distributional differences between sequences of images

    $\mathbf{S}_{<t}=\{\mathbf{X}_{t-T}, \mathbf{X}_{t-T+1}, \ldots, \mathbf{X}_t\}$ preceding an event ($Y_t=1$) vs non-event ($Y_t=0$).
  - The problem is difficult because **the data are high-dimensional**.
  - The data $\{(\mathbf{S}_{<t}, Y_t)\}_{t\geq 0}$ are also **not IID** because of strong temporal dependence.

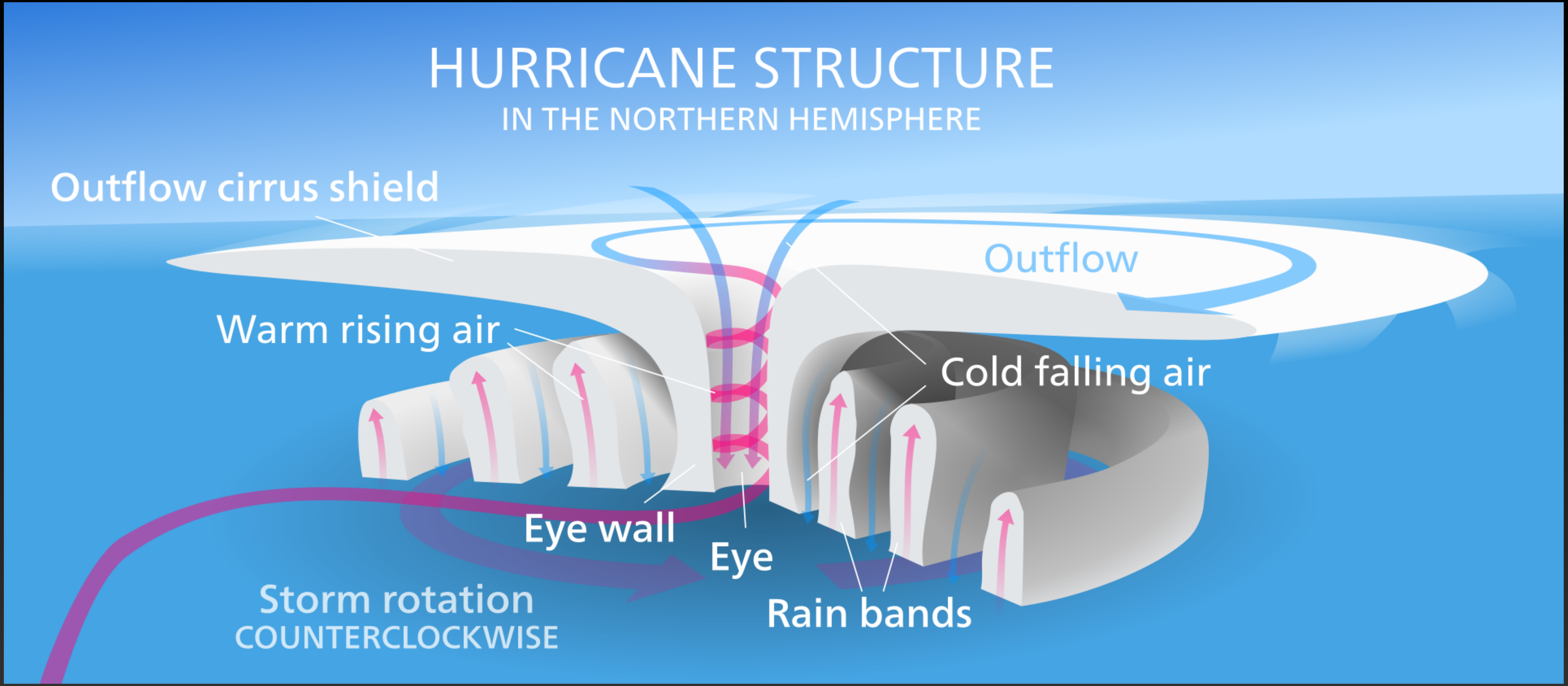
Tropical Cyclones (TCs) are Rapidly Rotating Systems

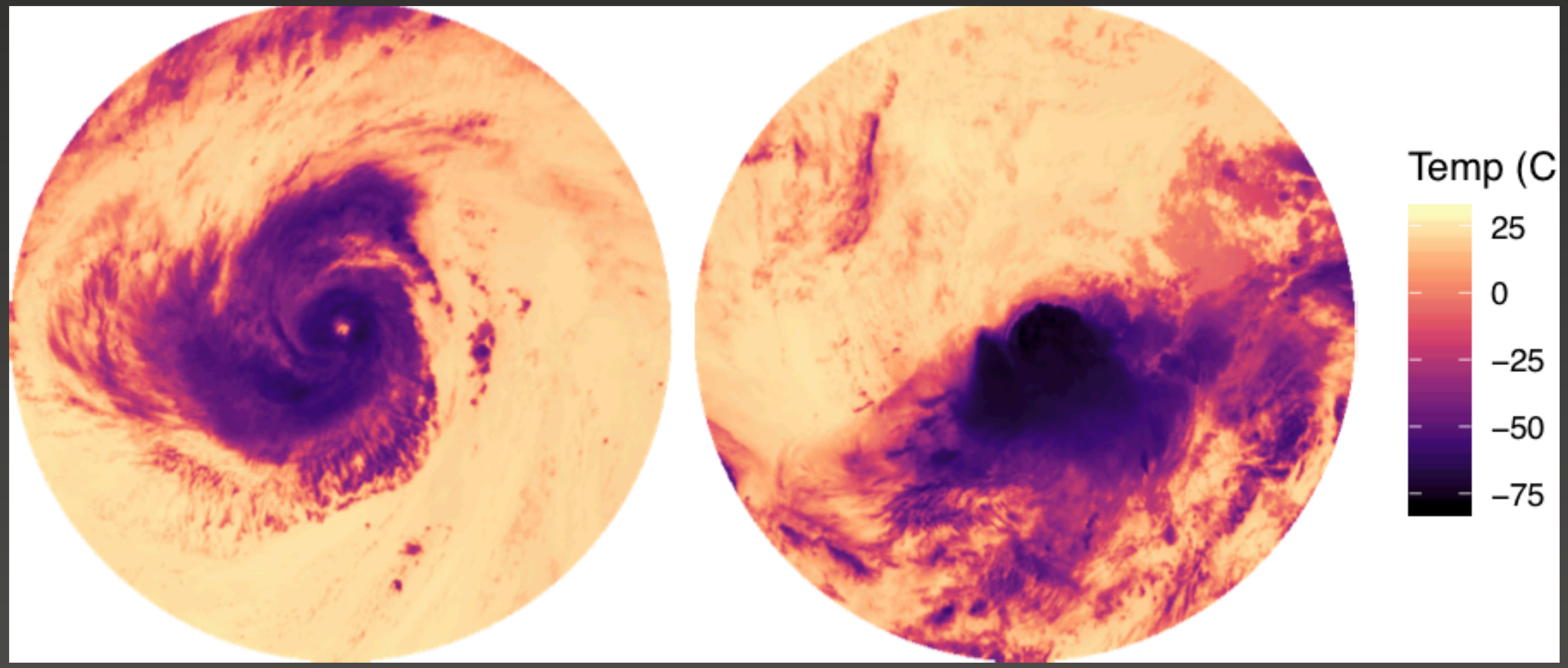
Develop over Warm Tropical Waters





- Because TCs develop far from land-based observing networks, geostationary satellite imagery (GOES) is critical to monitor these storms.

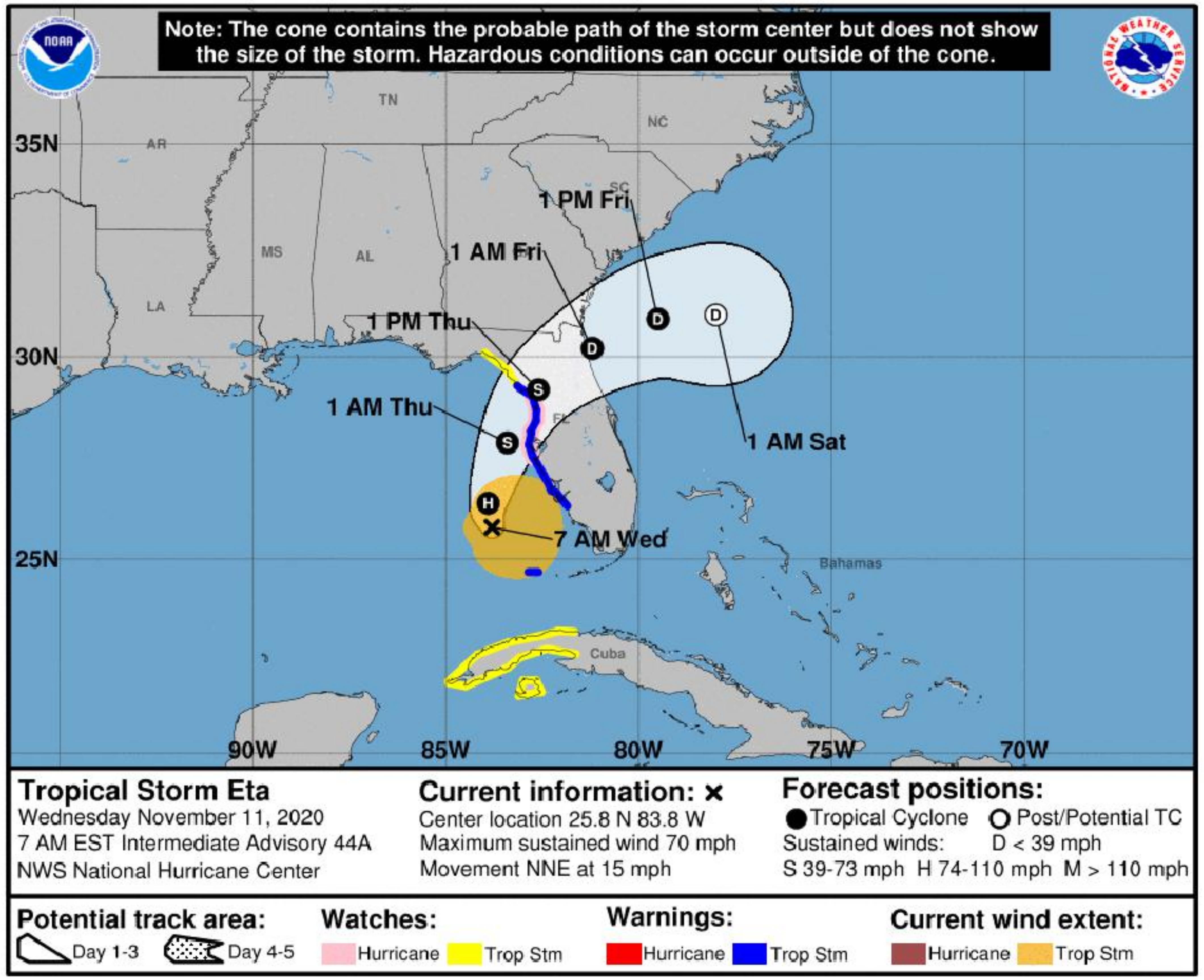


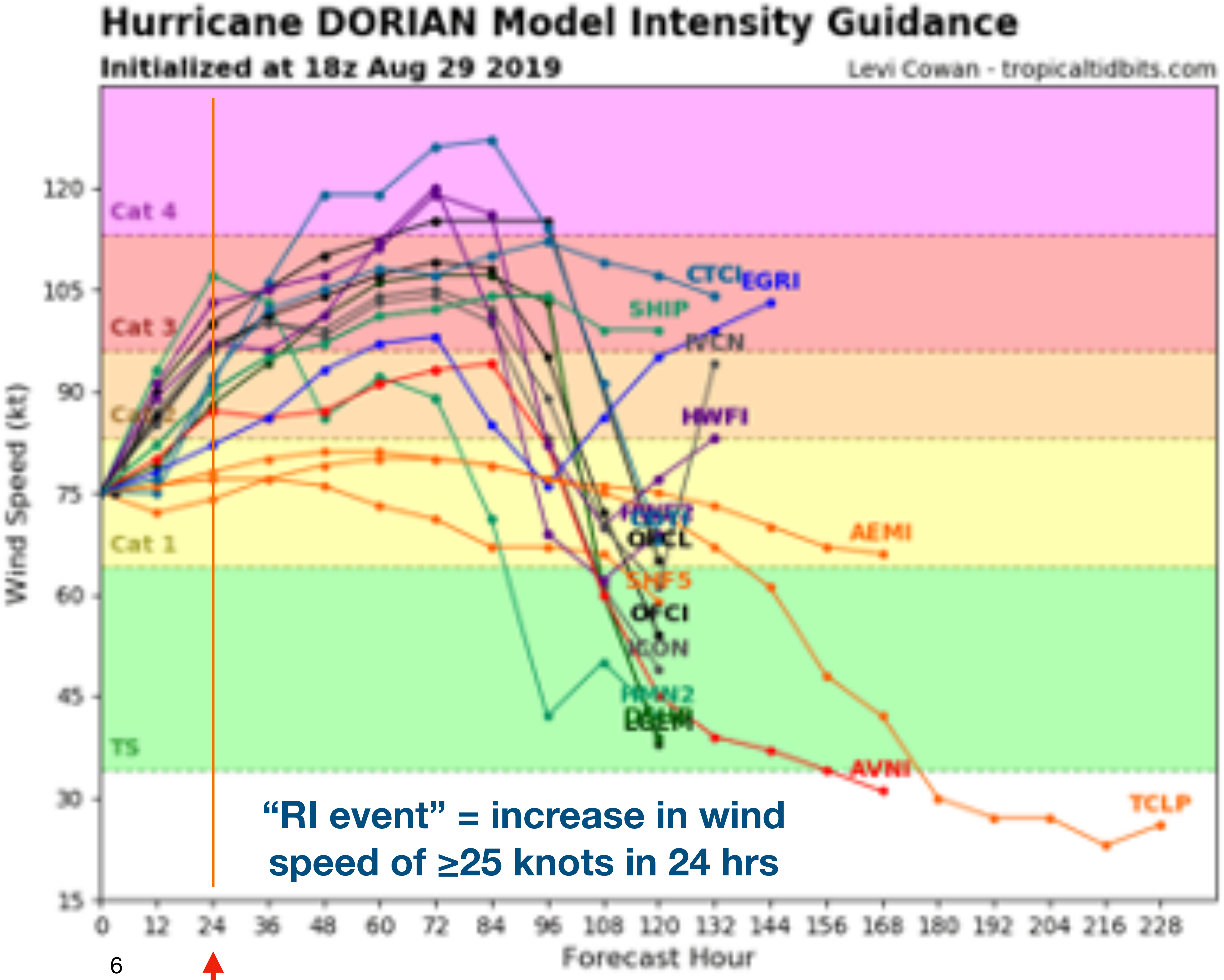


Left: Edouard 2014 (95 kt; Category 2); Right: Nicole 2016 (47 kt; TS)

Spatio-Temporal Information in IR Imagery Underutilized

Trajectory Forecasts vs. TC Short-term Intensity Forecasts (24-hr)



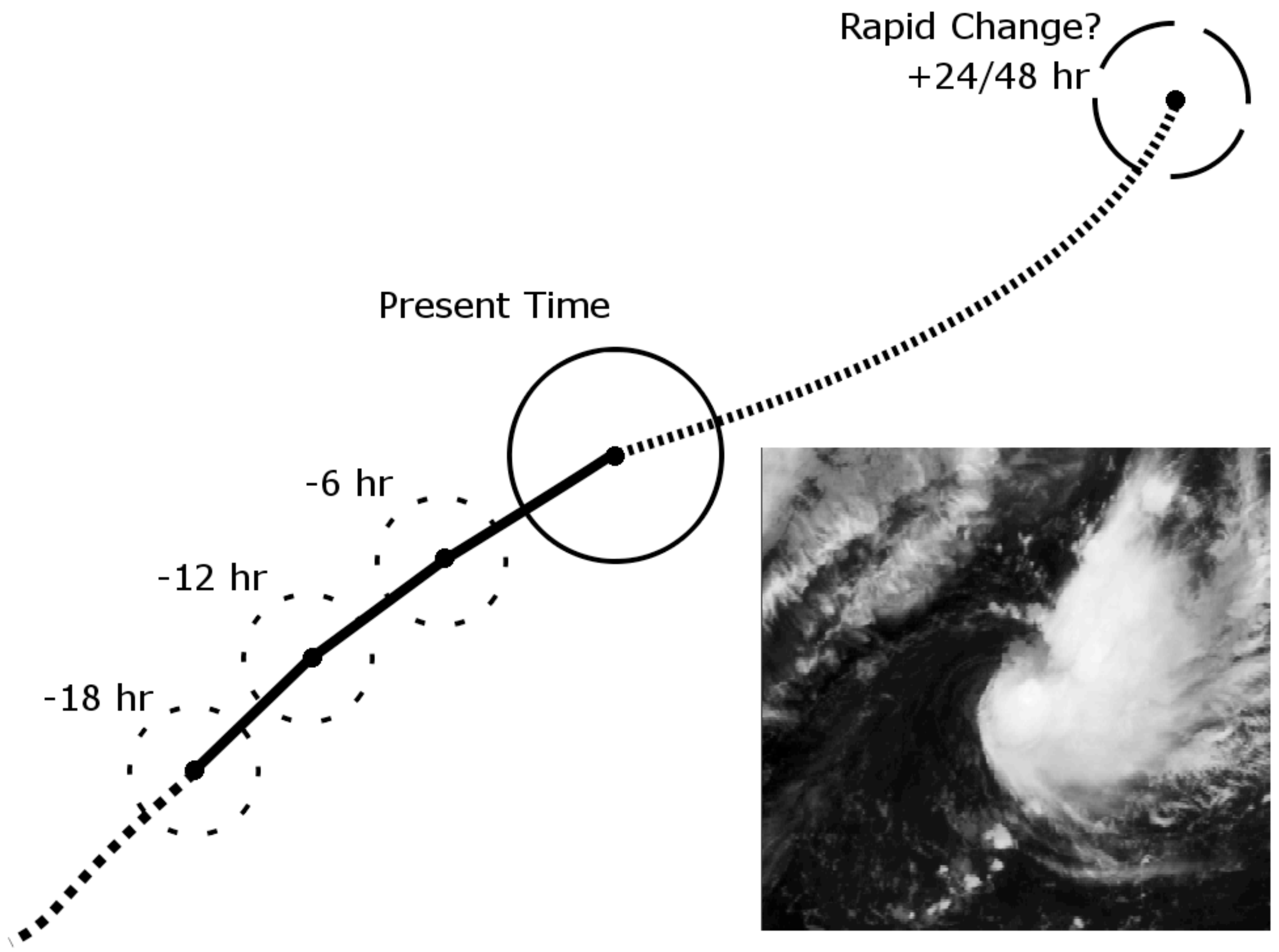


“RI event” = increase in wind speed of ≥25 knots in 24 hrs

Two Databases

TC location & intensity

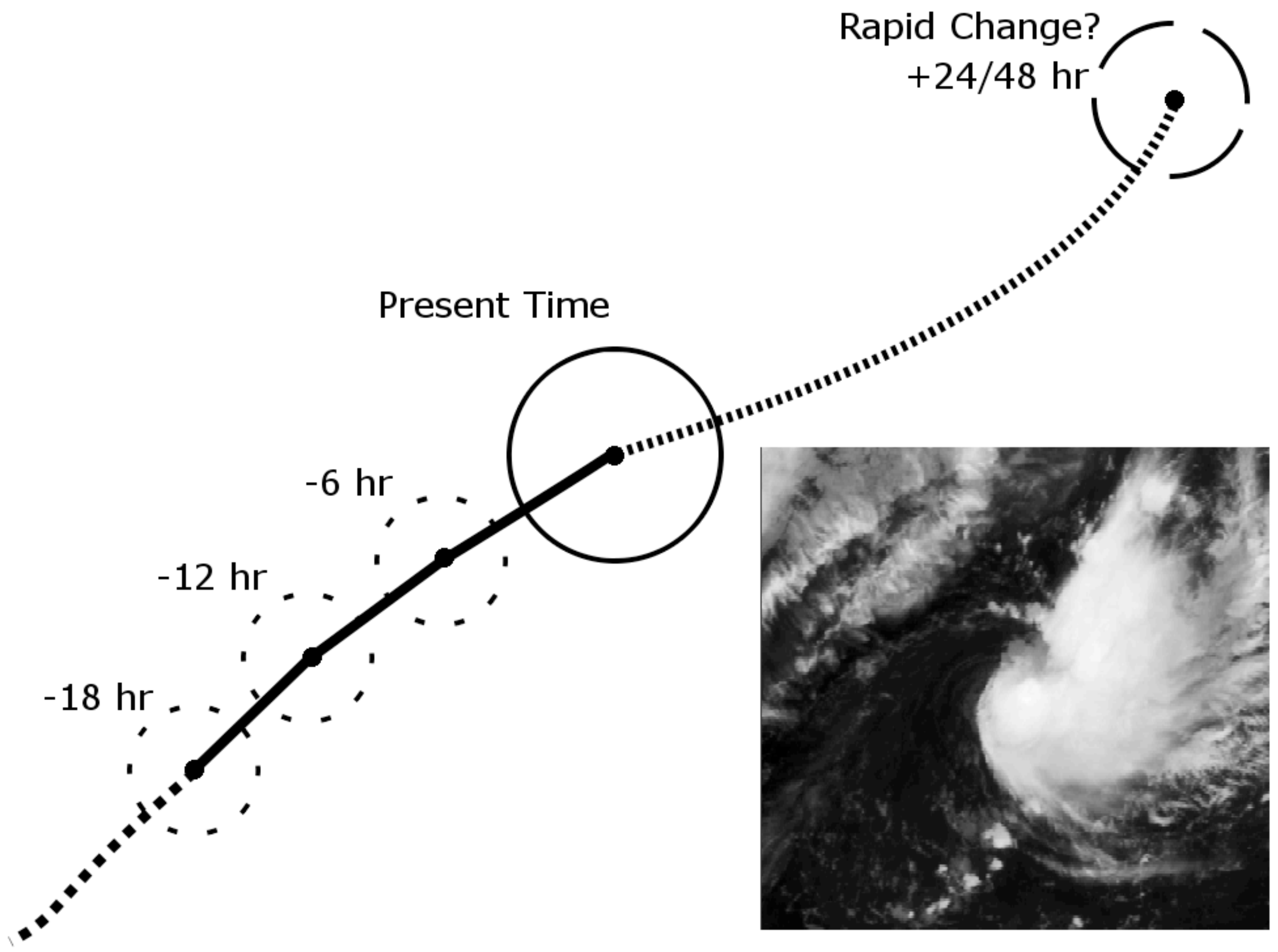
1. HURDAT2
   - Hurricane best-track data
   - 6-hr resolution (1979-2020)
   - TC location, intensity

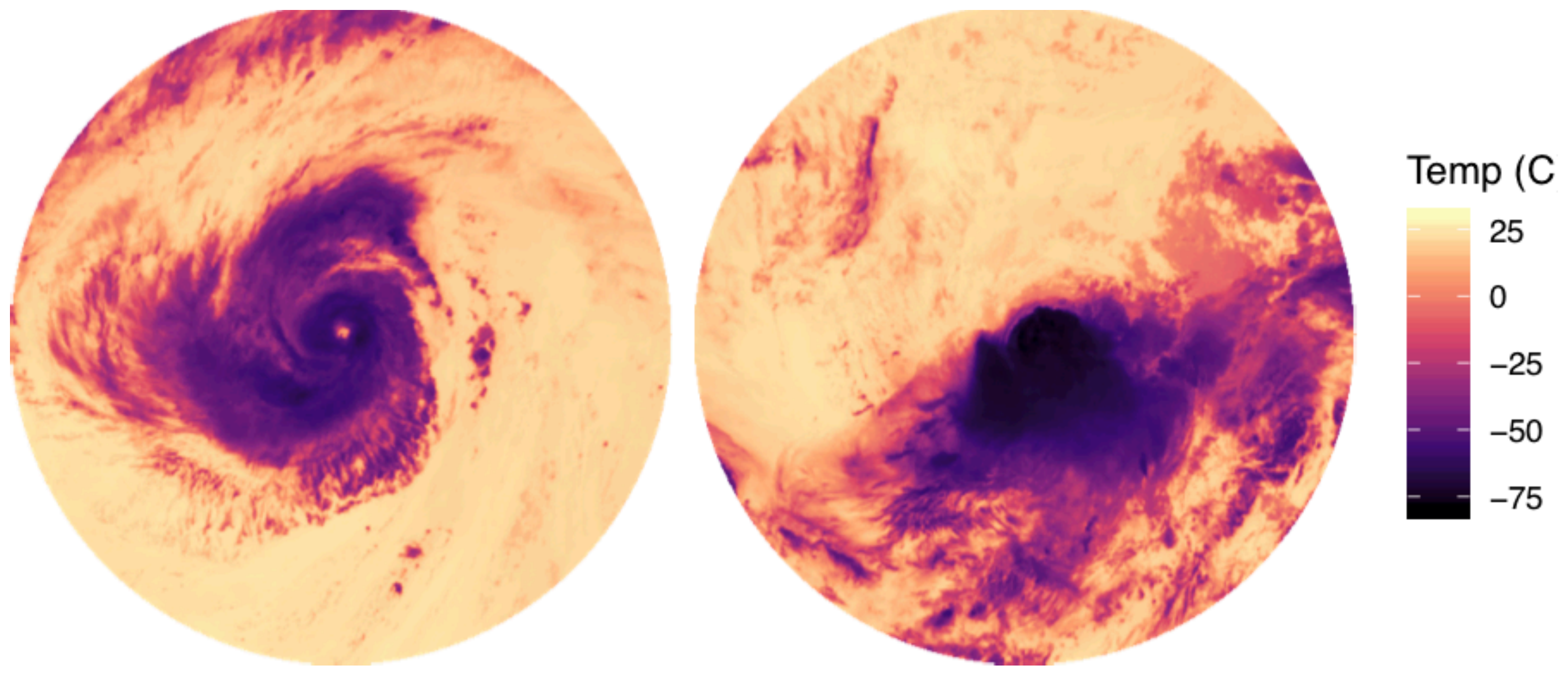


Two Databases

TC location & intensity + GOES images

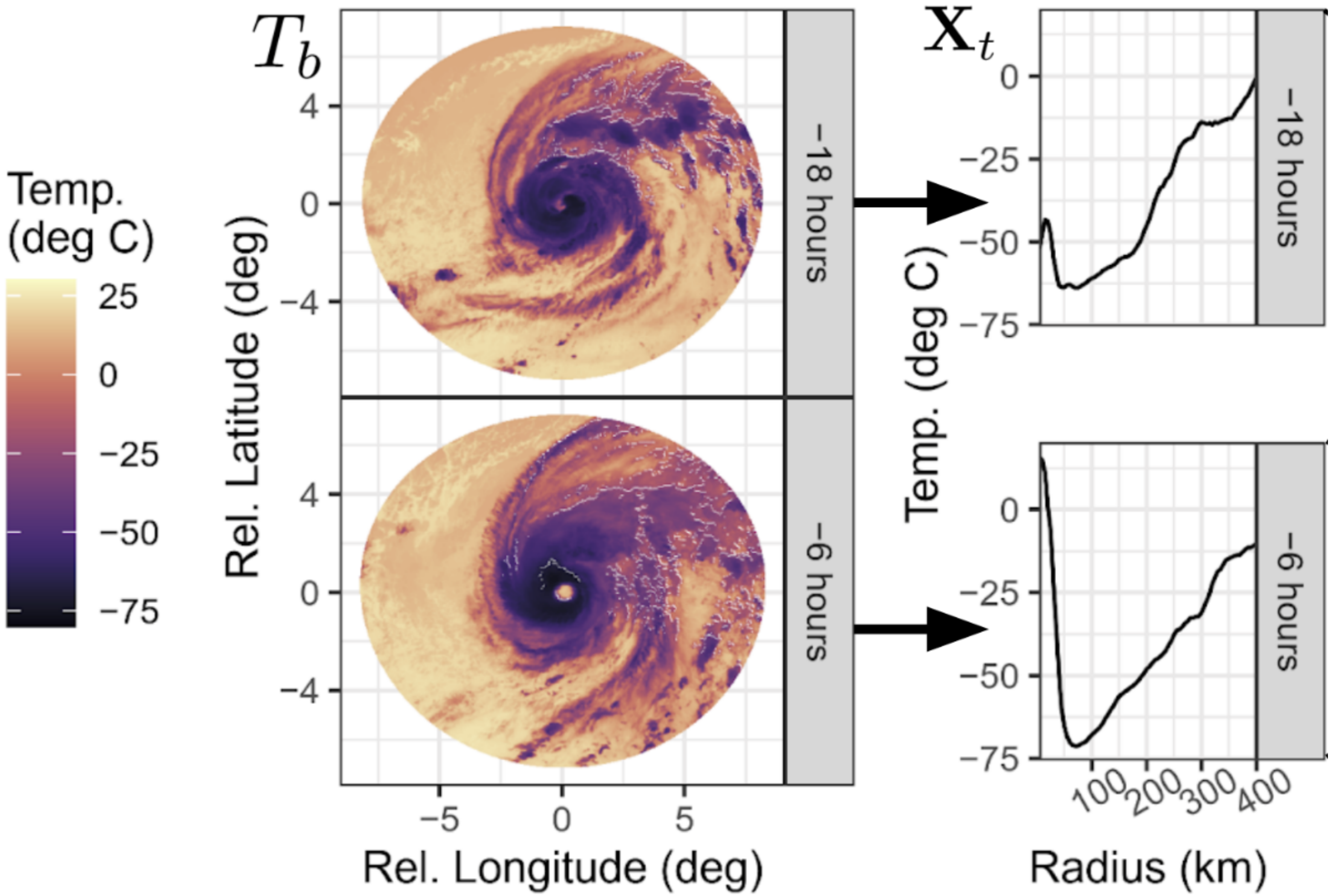
1. HURDAT2
   - Hurricane best-track data
   - 6-hr resolution (1979-2020)
   - TC location, intensity
2. MERGIR
   - Geostationary satellite (GOES) imagery
   - 4-km, 30-min resolution
   - 2000-2020





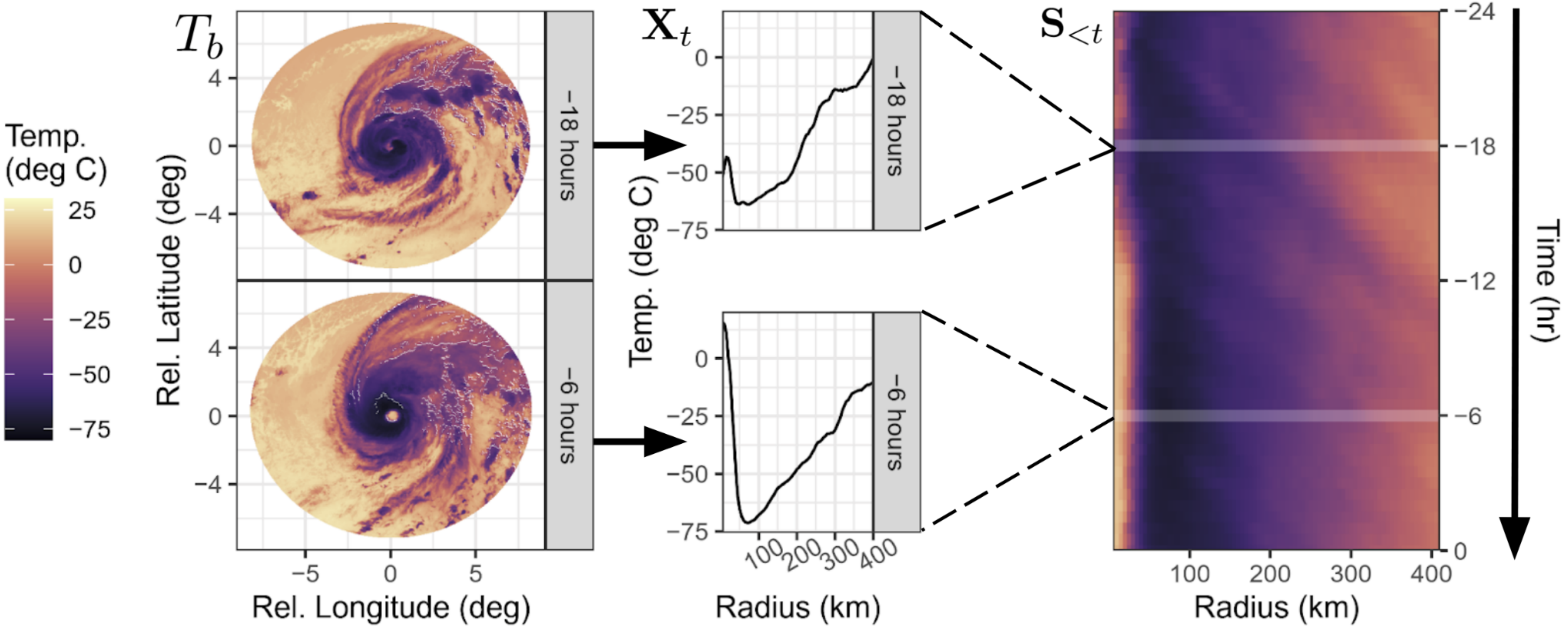
Evolution of TC Convective Structure

as "Structural Trajectories" $S_{<t}$ of Interpretable Functions X_t



Evolution of TC Convective Structure

as "Structural Trajectories" $S_{<t}$ of Interpretable Functions X_t



Structural trajectory is a 24h sequence of cont. functions (at 30 min time res). "Hovmöller diagram"

$$\mathbf{S}_{<t} = \{\mathbf{X}_{t-24h}, \mathbf{X}_{t-23.5h}, \mathbf{X}_{t-23h}, \ldots, \mathbf{X}_t\}$$

Main Questions as a Two-Sample Testing Problem

$$Y_t = \begin{cases} 1 & \text{if RI event at time } t, \\ 0 & \text{otherwise} \end{cases}$$

$$H_0 : p(\mathbf{s}_{<t}|Y_t = 1) = p(\mathbf{s}_{<t}|Y_t = 0) \text{ for all } \mathbf{s}_{<t} \in \mathcal{S}, \text{ versus}$$
$$H_1 : p(\mathbf{s}_{<t}|Y_t = 1) \neq p(\mathbf{s}_{<t}|Y_t = 0) \text{ for some } \mathbf{s}_{<t} \in \mathcal{S}.$$

- Does the distribution of structural trajectories differ between the lead-up to a RI vs. non-RI event? (Statistical significance)
- If there is a difference between the distributions, how do they differ? (Scientific interpretability)

Main Questions as a Two-Sample Testing Problem

$$Y_t = \begin{cases} 1 & \text{if RI event at time } t, \\ 0 & \text{otherwise} \end{cases}$$

$$H_0 : p(\mathbf{s}_{<t}|Y_t = 1) = p(\mathbf{s}_{<t}|Y_t = 0) \text{ for all } \mathbf{s}_{<t} \in \mathcal{S}, \text{ versus}$$
$$H_1 : p(\mathbf{s}_{<t}|Y_t = 1) \neq p(\mathbf{s}_{<t}|Y_t = 0) \text{ for some } \mathbf{s}_{<t} \in \mathcal{S}.$$

- Does the distribution of structural trajectories differ between the lead-up to a RI vs. non-RI event? (Statistical significance)
- If there is a difference between the distributions, how do they differ? (Scientific interpretability)

Why the Two-Sample Test is Challenging …

$$H_0 : p(\mathbf{s}_{<t}|Y_t = 1) = p(\mathbf{s}_{<t}|Y_t = 0) \text{ for all } \mathbf{s}_{<t} \in \mathcal{S}, \text{ versus}$$
$$H_1 : p(\mathbf{s}_{<t}|Y_t = 1) \neq p(\mathbf{s}_{<t}|Y_t = 0) \text{ for some } \mathbf{s}_{<t} \in \mathcal{S}.$$

- The complexity of the data itself, with *one observation* representing an entire sequence $\mathbf{S}_{<t}$ of functions

$$\mathbf{S}_{<t} = \{\mathbf{X}_{t-T}, \mathbf{X}_{t-T+1}, \ldots, \mathbf{X}_t\}$$

- Dependence between labels Y_t (and sequences $\mathbf{S}_{<t}$) at nearby time points t
  - IID data ⇒ ``Dependent Identically Distributed'' (DID) sequence data

$$\{(\mathbf{S}_{<t}, Y_t)\}_{t \geq 0}$$

Two-Sample Test via Regression (HighDim IID data)

[Freeman, Kim & Lee, MNRAS 2017; Kim, Lee & Lei, EJS 2019]

Suppose we have two samples:

$$\mathbf{S}_1^0, \ldots, \mathbf{S}_{n_0}^0 \sim P_0 \quad \text{and} \quad \mathbf{S}_1^1, \ldots, \mathbf{S}_{n_1}^1 \sim P_1$$

A two sample-test would ask whether P_0 and P_1 are the same; i.e., it would test the null hypothesis

$$H_0 : p(\mathbf{s}|Y = 0) = p(\mathbf{s}|Y = 1) \text{ for all } \mathbf{s} \in \mathcal{S}$$

Two-Sample Test via Regression (HighDim IID data)

[Freeman, Kim & Lee, MNRAS 2017; Kim, Lee & Lei, EJS 2019]

Suppose we have two samples:

$$\mathbf{S}_1^0, \ldots, \mathbf{S}_{n_0}^0 \sim P_0 \quad \text{and} \quad \mathbf{S}_1^1, \ldots, \mathbf{S}_{n_1}^1 \sim P_1$$

A two sample-test would ask whether P_0 and P_1 are the same; i.e., it would test the null hypothesis

$$H_0 : p(\mathbf{s}|Y = 0) = p(\mathbf{s}|Y = 1) \text{ for all } \mathbf{s} \in \mathcal{S}$$

↕

By Bayes rule, this is equivalent to testing

$$H_0 : \mathbb{P}(Y = 1|\mathbf{S} = \mathbf{s}) = \mathbb{P}(Y = 1) \text{ for all } \mathbf{s} \in \mathcal{S}$$

Convert 2-sample testing to a regression problem

Our null and alternative hypotheses are

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

Define the regression function $m_{\text{post}}(\mathbf{s}) := \mathbb{P}(Y = 1|\mathbf{S} = \mathbf{s})$.
Let $\widehat{m}(\mathbf{s})$ be an estimate of $m_{\text{post}}(\mathbf{s})$ based on train data $\mathcal{T} = \{(\mathbf{S}_i, Y_i)\}_{i=1}^n$.
Let $\widehat{m}_{\text{prior}}(\mathbf{s}) = \frac{1}{n}\sum_{i=1}^n I(Y_i = 1)$ be an estimate of $m_{\text{prior}} := \mathbb{P}(Y = 1)$.

Convert 2-sample testing to a regression problem

Our null and alternative hypotheses are

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

Define the regression function $m_{\text{post}}(\mathbf{s}) := \mathbb{P}(Y=1|\mathbf{S}=\mathbf{s})$.
Let $\widehat{m}(\mathbf{s})$ be an estimate of $m_{\text{post}}(\mathbf{s})$ based on train data $\mathcal{T} = \{(\mathbf{S}_i, Y_i)\}_{i=1}^{n}$.
Let $\widehat{m}_{\text{prior}}(\mathbf{s}) = \frac{1}{n}\sum_{i=1}^{n} I(Y_i = 1)$ be an estimate of $m_{\text{prior}} := \mathbb{P}(Y=1)$.

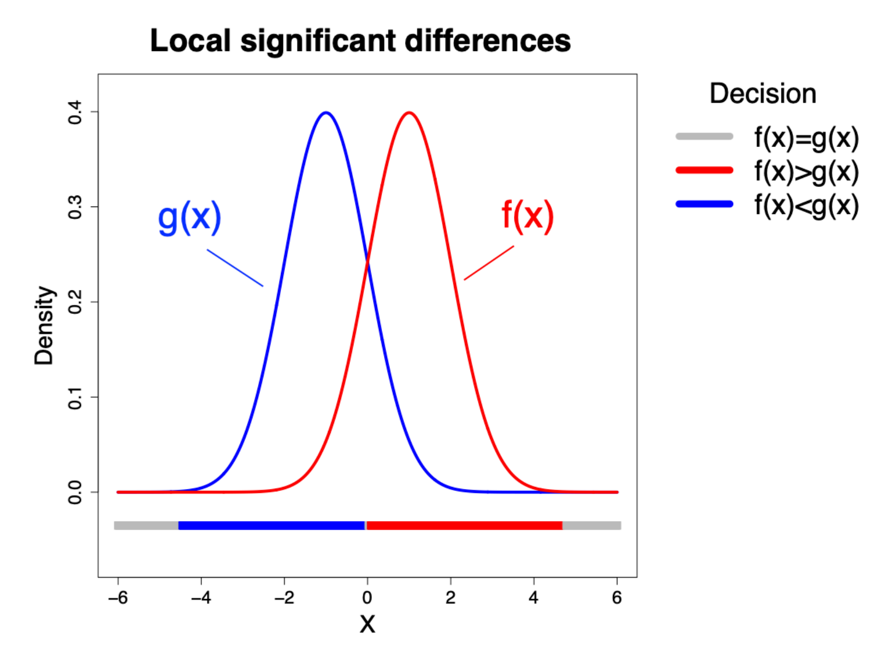
Define the "local posterior difference" (LPD) at evaluation points $\mathcal{V} \subset \mathcal{S}$:

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

Our global test statistic is

$$\lambda := \frac{1}{|\mathcal{V}|} \sum_{\mathbf{s} \in \mathcal{V}} \lambda(\mathbf{s})^2$$

Can Detect Distributional Differences in Galaxy Images for HighSF and LowSF Samples [Freeman, Kim & Lee, MNRAS 2017]



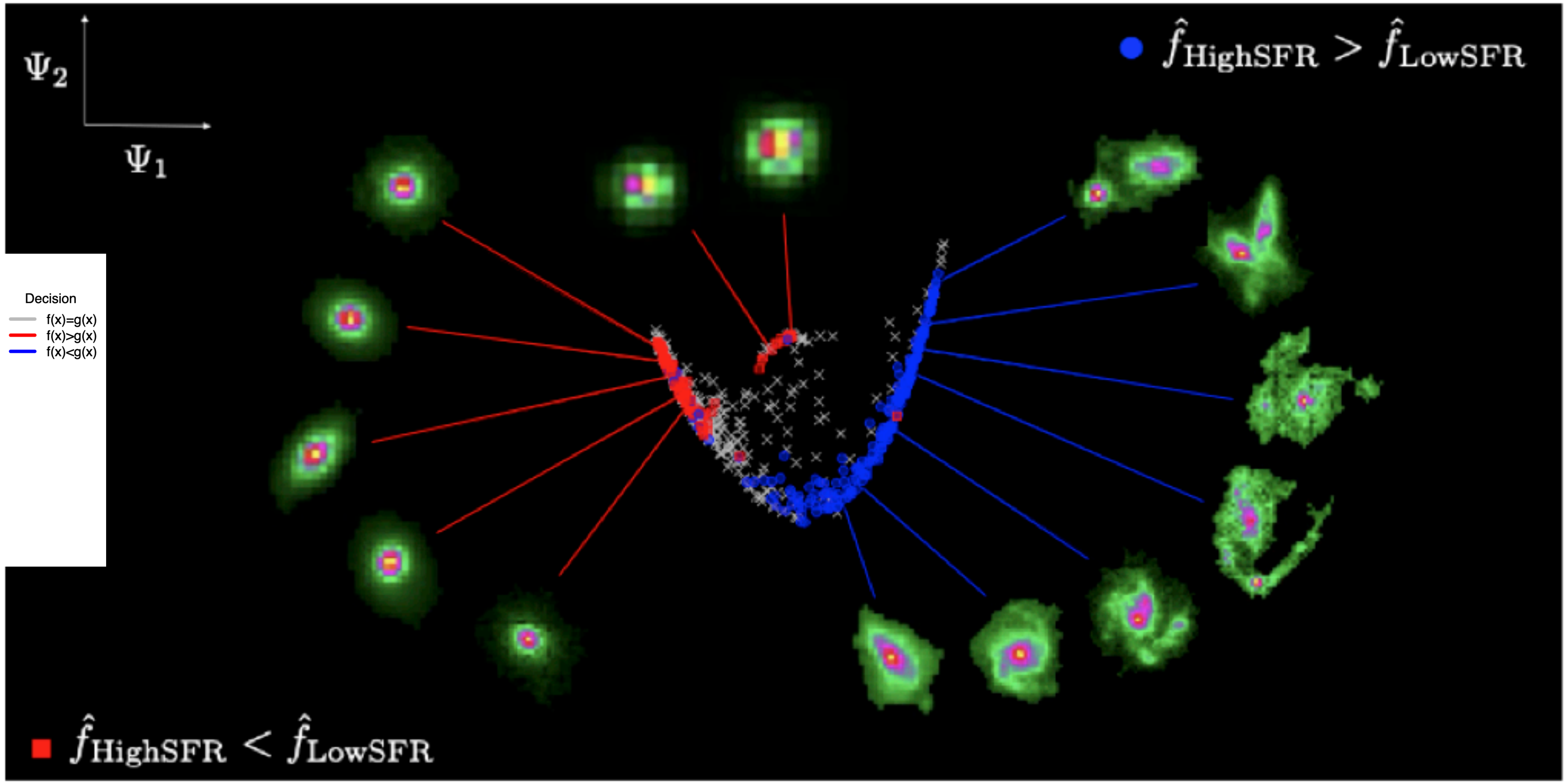
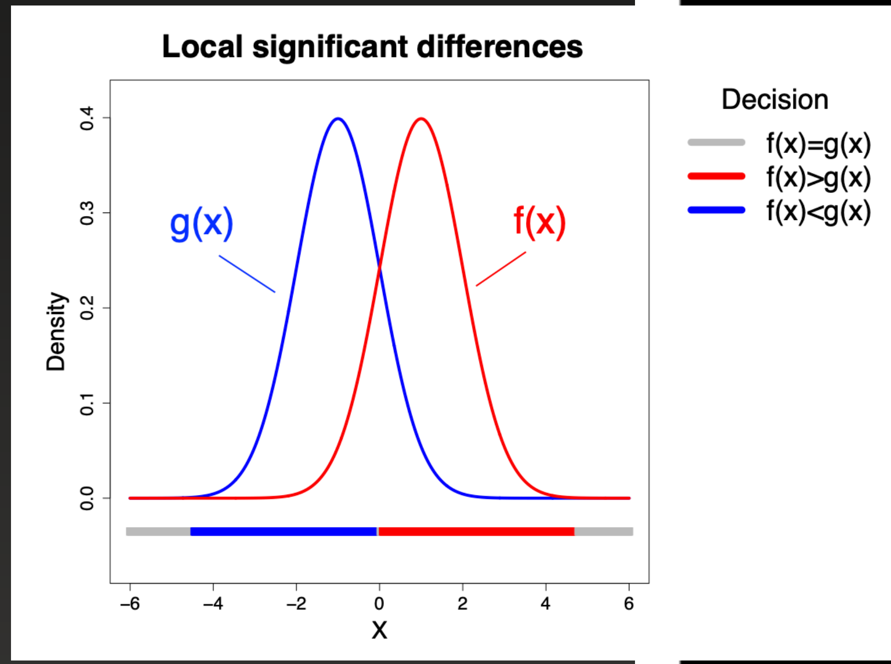


Figure 9: Results of two-sample testing of point-wise differences between high- and low-SFR galaxies in a seven-dimensional morphology space. The red color indicates regions where the density of low-SFR galaxies are significantly higher, and the blue color indicates regions that are dominated by high-SFR galaxies. The test points are visualized via a two-dimensional diffusion map. Figure adapted from [49].

Can Detect Distributional Differences in Galaxy Images for HighSF and LowSF Samples [Freeman, Kim & Lee, MNRAS 2017]



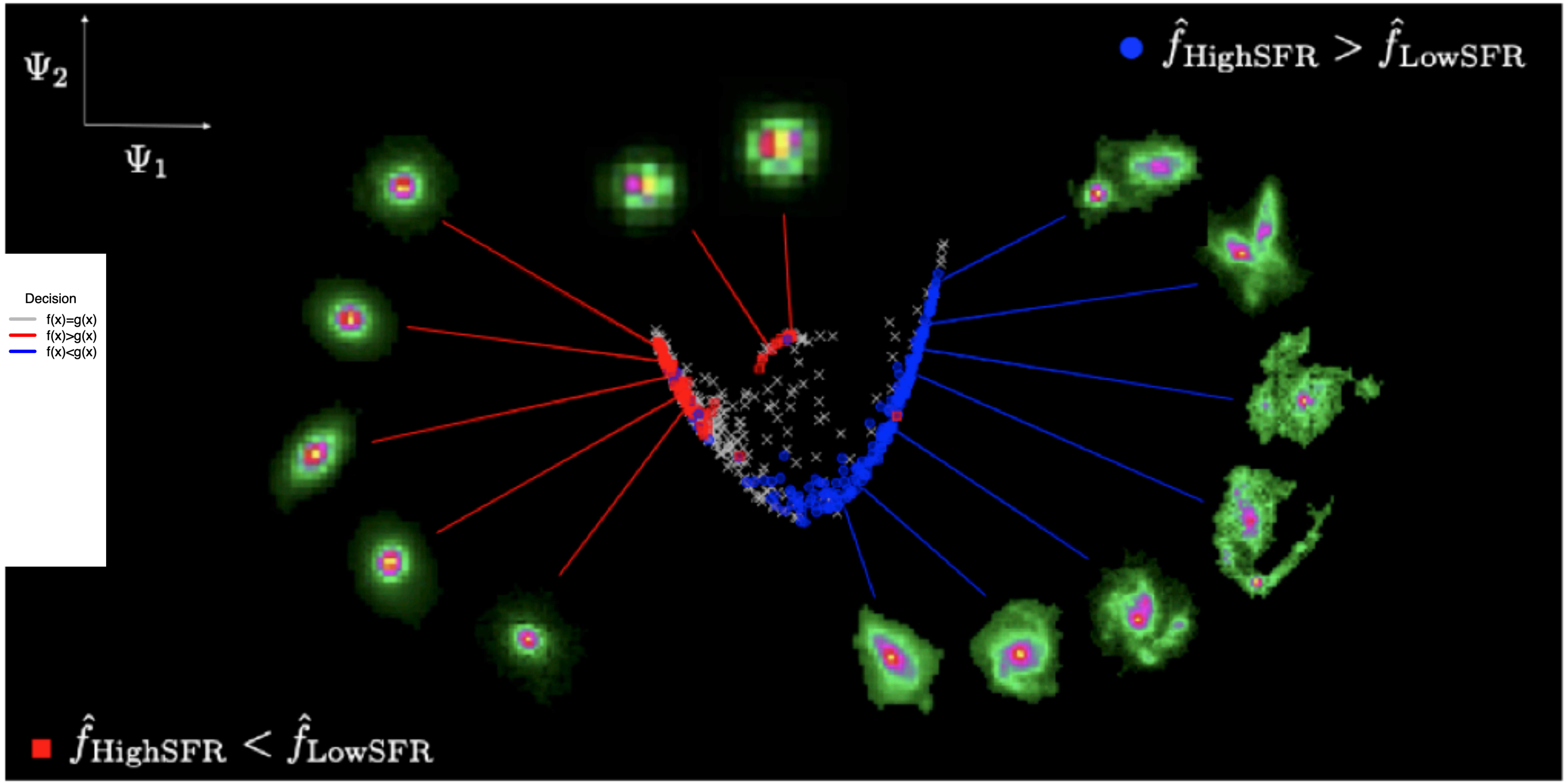
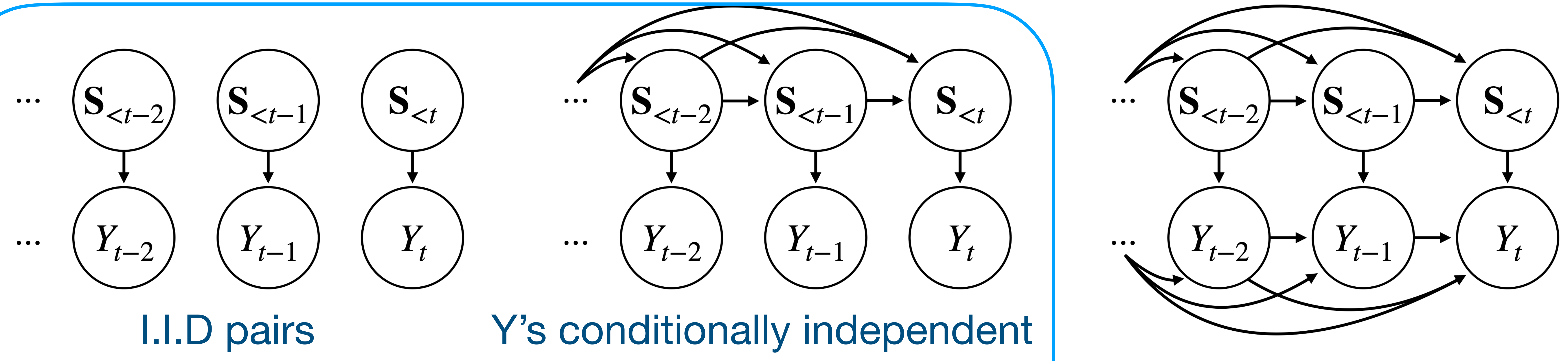


Figure 9: Results of two-sample testing of point-wise differences between high- and low-SFR galaxies in a seven-dimensional morphology space. The red color indicates regions where the density of low-SFR galaxies are significantly higher, and the blue color indicates regions that are dominated by high-SFR galaxies. The test points are visualized via a two-dimensional diffusion map. Figure adapted from [49].

But these are I.I.D data and not dependent sequence data…

Dependence Settings for Labeled Sequence Data

$$\mathbf{S}_{<t} = \{\mathbf{X}_{t-T}, \mathbf{X}_{t-T+1}, \ldots, \mathbf{X}_t\}$$



(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.

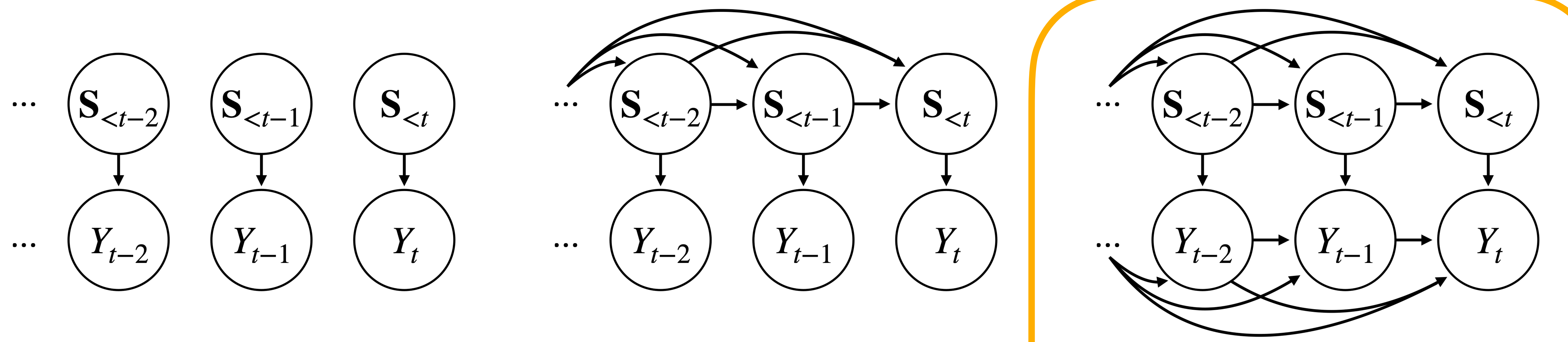
(c) Setting C: Y_t conditionally dependent on Y_{t-1} given $\mathbf{S}_{<t}$; $\mathbf{S}_{<t}$ and Y_t are each autocorrelated.

- In Settings A and B: Labels Y are conditionally independent given S

⇒ Labels Y are exchangeable under H_0. A permutation test would be valid [Kim et al 2019]

TC Data are Not Exchangeable.

$$\mathbf{S}_{<t} = \{\mathbf{X}_{t-T}, \mathbf{X}_{t-T+1}, \ldots, \mathbf{X}_t\}$$



(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.

- In TC data, we have auto-correlation in Y which is inherent or governed by unobserved quantities (Setting C) ⇒ Permutation tests are not valid.

Permutation Test ⇒ Markov Chain (MC) Bootstrap

For permutation test:

- Estimate $m_{\text{post}}(\mathbf{s}) := \mathbb{P}(Y_t = 1 | \mathbf{S}_{<t} = \mathbf{s})$ using labeled train data $\{(\mathbf{S}_{<t}, Y_t)\}_{t \in \mathcal{T}_1}$
- Compute test statistic $\lambda = \sum_{\mathbf{s} \in \mathcal{V}} \lambda^2(\mathbf{s}),$ where $\lambda(\mathbf{s}) = \widehat{m}_{\text{post}}(\mathbf{s}) - \widehat{m}_{\text{prior}}$
- To estimate the null distribution of λ:
  - Permute original labels $\{Y_t\}_{t \in \mathcal{T}_1}$
  - Recompute test statistic λ

Permutation Test $\Rightarrow$ Markov Chain (MC) Bootstrap Test

For permutation test:

- Estimate $m_{\text{post}}(\mathbf{s}) := \mathbb{P}(Y_t = 1|\mathbf{S}_{<t} = \mathbf{s})$ using labeled train data $\{(\mathbf{S}_{<t}, Y_t)\}_{t\in\mathcal{T}_1}$
- Compute test statistic $\lambda = \sum_{\mathbf{s}\in\mathcal{V}} \lambda^2(\mathbf{s})$, where $\lambda(\mathbf{s}) = \widehat{m}_{\text{post}}(\mathbf{s}) - \widehat{m}_{\text{prior}}$
- To estimate the null distribution of λ:
  - Permute original labels $\{Y_t\}_{t\in\mathcal{T}_1}$
  - Recompute test statistic λ

Instead, use train data $\{Y_t\}_{t\in\mathcal{T}_2}$ and regression method to estimate

$$m_{\text{seq}}(Y_{t-1}, \ldots, Y_{t-k}) := \mathbb{P}(Y_t = 1|Y_{t-1}, \ldots, Y_{t-k})$$

Draw new labels

$$\widetilde{Y}_t \sim \text{Binom}(\widehat{\mathbb{P}}(Y_t = 1|Y_{t-1}, \ldots, Y_{t-k})) \quad \text{for} \quad t \in \mathcal{T}_1$$

TC train data: High-res GOES images back to 2000 (~400 TCs to fit regression of Y on S). However, intensity data goes back to 1979 (>1000 TCs to fit MC on labels)

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**

Theorem: MC Bootstrap Test is Valid Asymptotically

Assume:

1. $\{(\mathbf{S}_{<t}, Y_t)\}_{t\geq 0}$ is a stationary sequence
2. $\{(\mathbf{S}_{<t}, Y_t)\}_{t\geq 0}$ satisfies the DAG of Setting C
3. $\widehat{m}_{\text{post}}$ is a continuous function of the train data $\mathcal{D} := \{(\mathbf{S}_{<t}, Y_t)\}_{t\in\mathcal{T}_1}$
4. the marginal distribution estimator is consistent; that is, the generator of $\{Y_t^0\}_{t\in\mathcal{T}_1}$ converges to the true generating process of $\{Y_t\}_{t\in\mathcal{T}_1}$ under H_0,

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

Theorem: MC Bootstrap Test is Valid Asymptotically

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

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

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

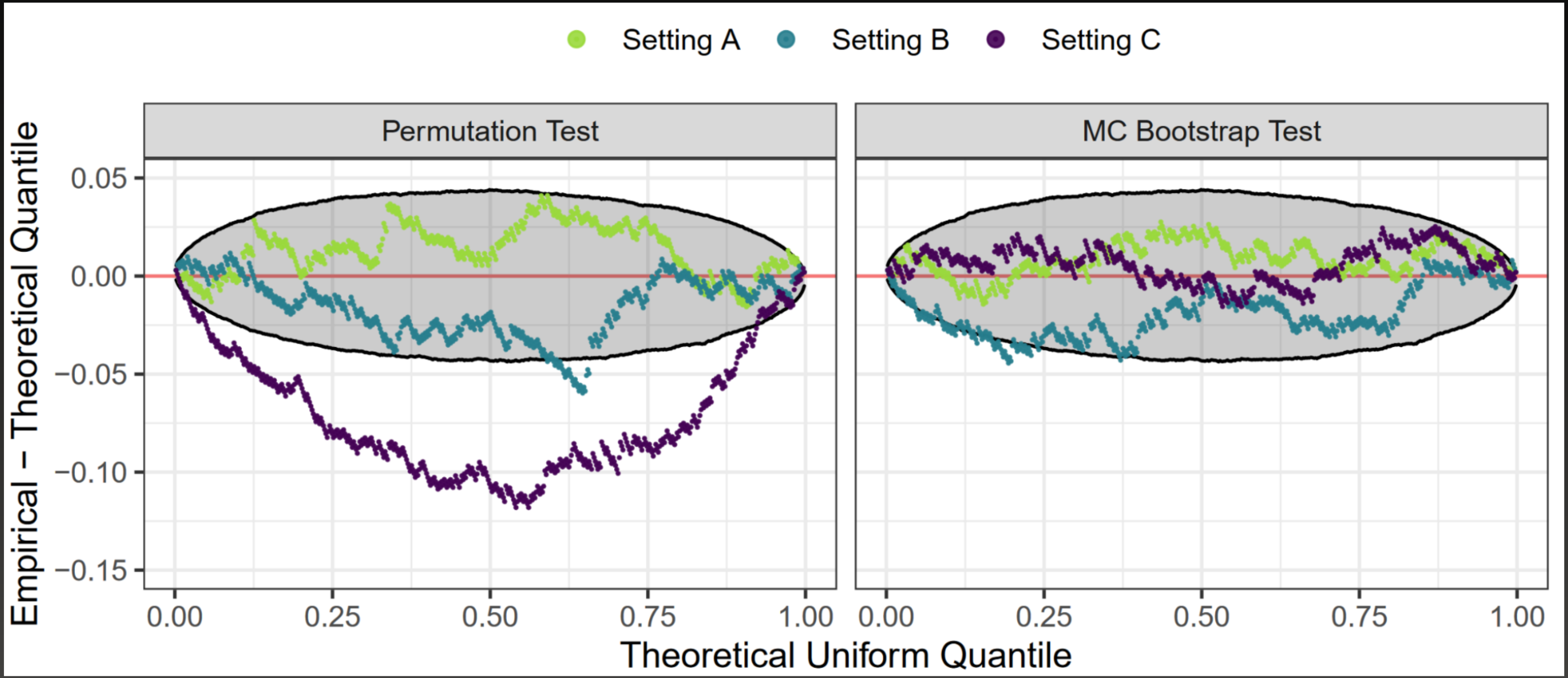
COROLLARY 1 (**Type I error control**). *Let*

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

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 $0 < \alpha < 1$,

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

Empirical Results for Synthetic Data Support Our Approach



(Left) Permutation test breaks under Setting C. (Right) MC bootstrap test still valid

<u>TC Analysis by Basin:</u> Reject $H_0: p(s_{<t}|Y_t=1)=p(s_{<t}|Y_t=0)$. Now what?

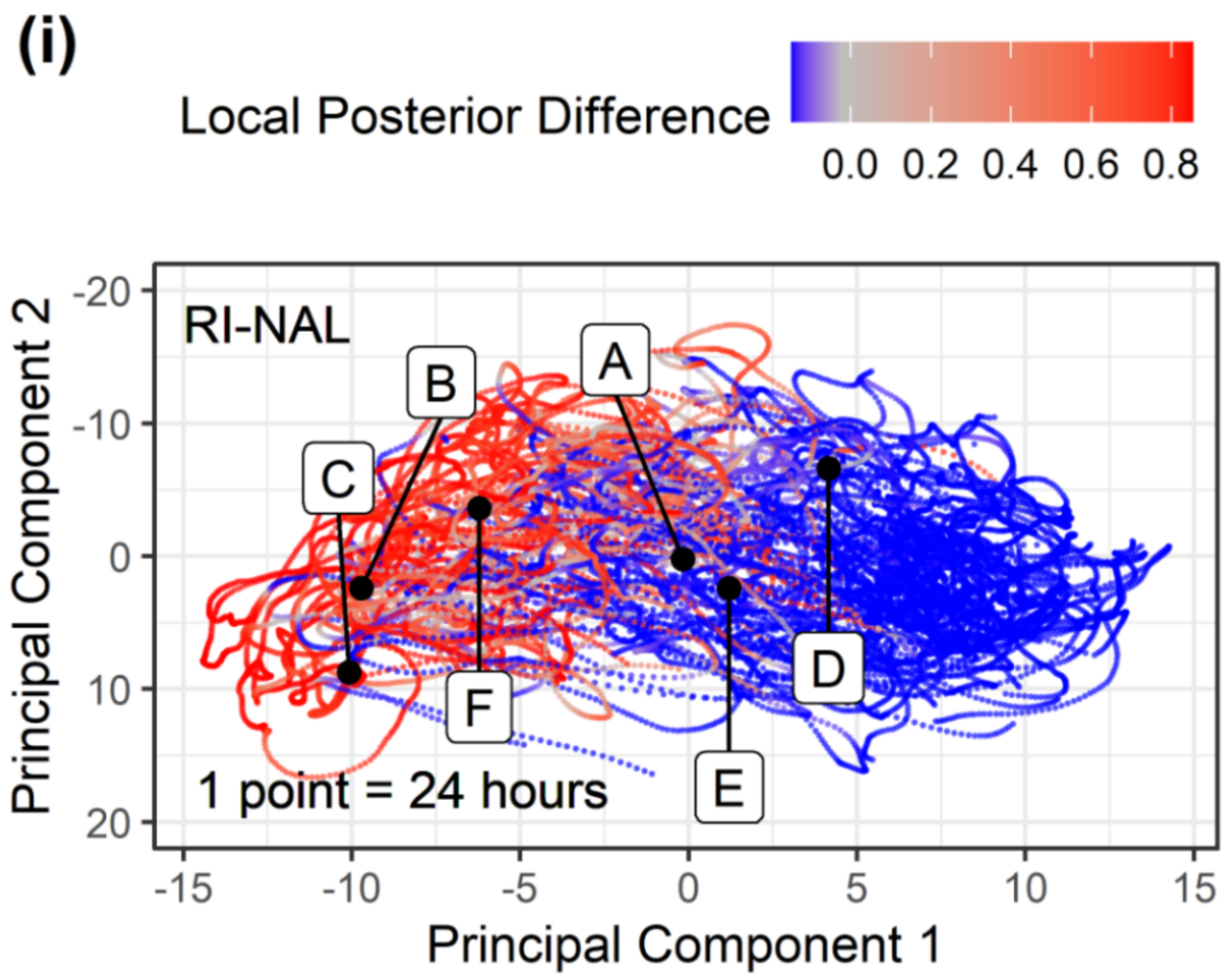
How **do the distributions of the structural trajectories $s_{<t}$ differ?**

<u>TC Analysis by Basin:</u> Reject H_0:$p(s_{<t}|Y_t=1)=p(s_{<t}|Y_t=0)$. Now what?

How do the distributions of the structural trajectories $s_{<t}$ differ?

- Use contributions to test statistic as a local diagnostic. "Local posterior difference" (LPD):

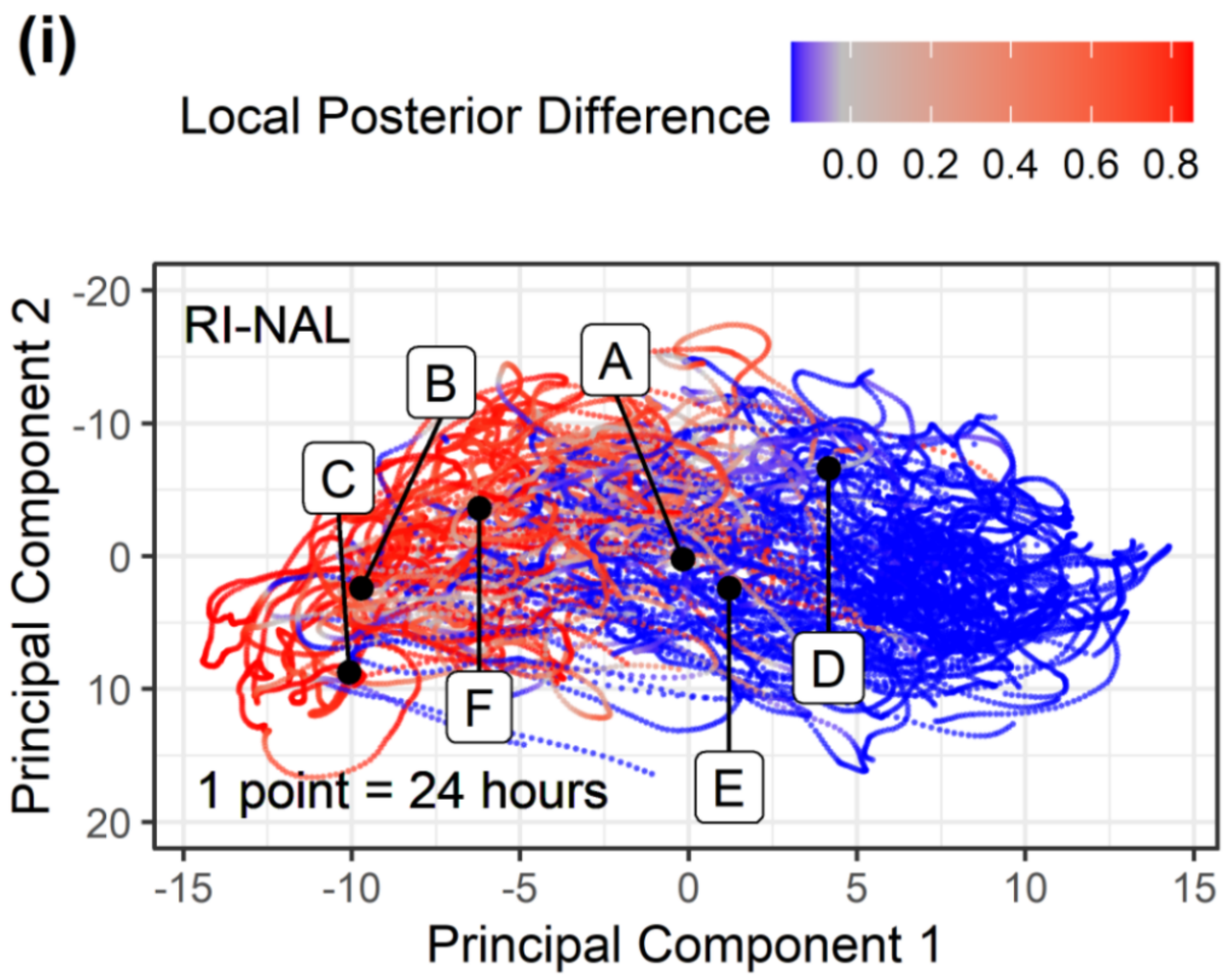
$$\lambda(\mathbf{s}) = \widehat{\mathbb{P}}(Y_t = 1|\mathbf{S}_{<t} = \mathbf{s}) - \widehat{\mathbb{P}}(Y_t = 1)$$

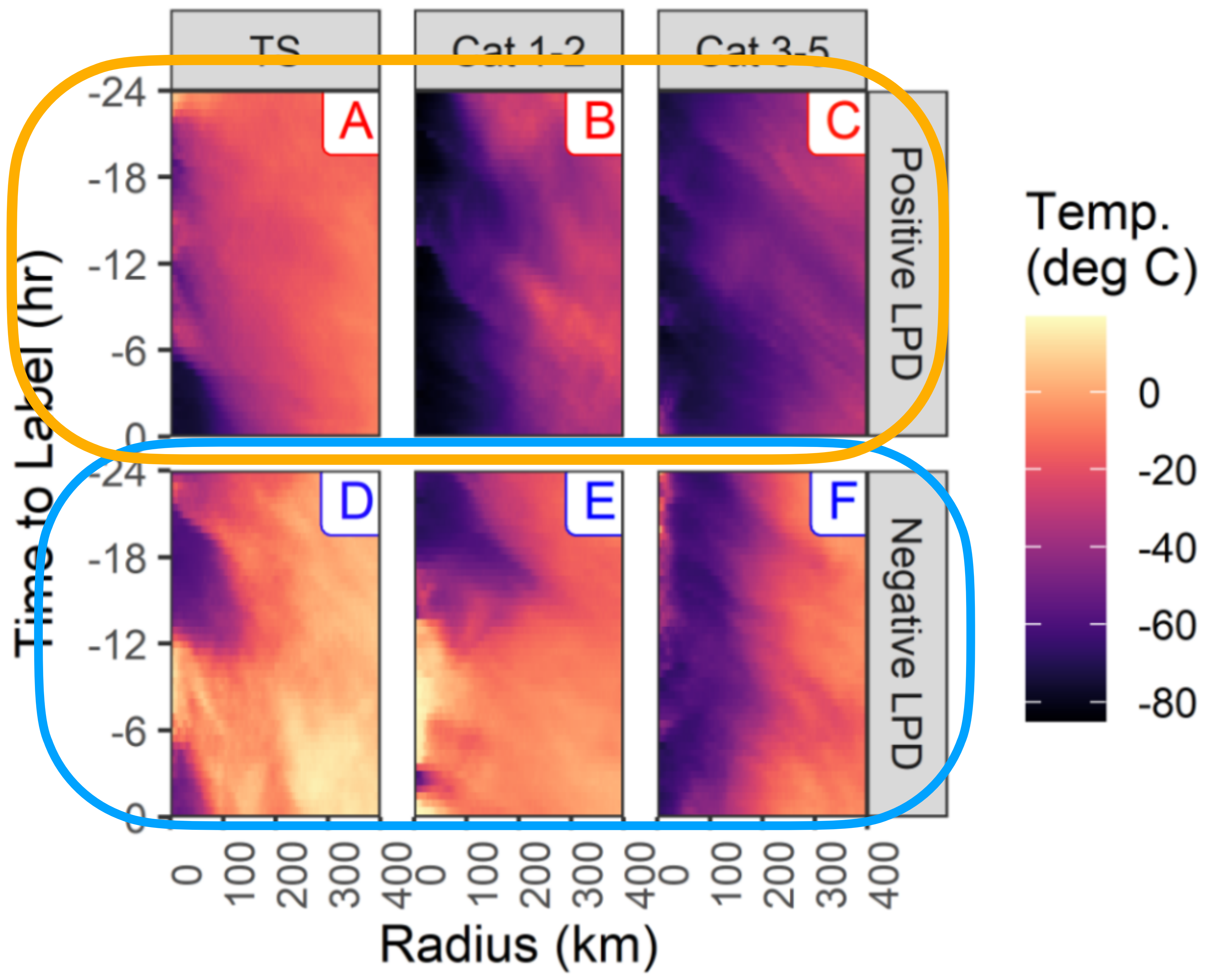


Positive LPD identifies trajectories with "high chance of RI"

Negative LPD identifies trajectories with "low chance of RI"

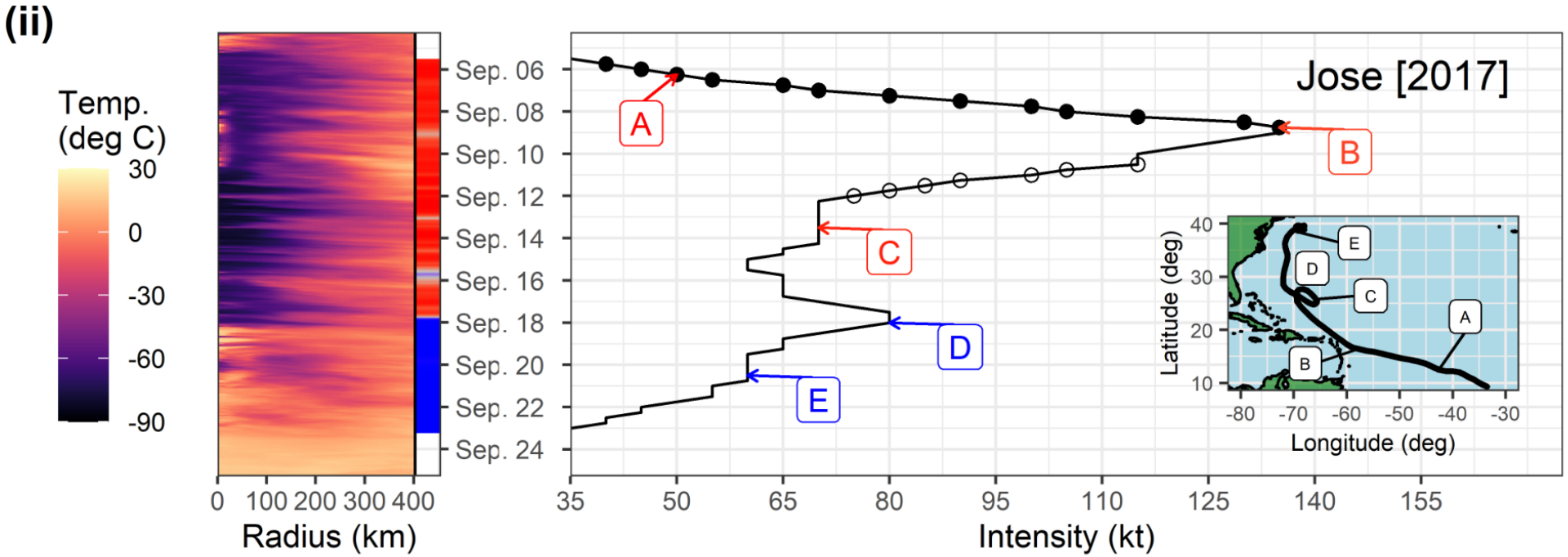
$$\lambda(\mathbf{s}) = \widehat{\mathbb{P}}(Y_t = 1 | \mathbf{S}_{<t} = \mathbf{s}) - \widehat{\mathbb{P}}(Y_t = 1)$$





LPDs can also be used to track development of specific TCs

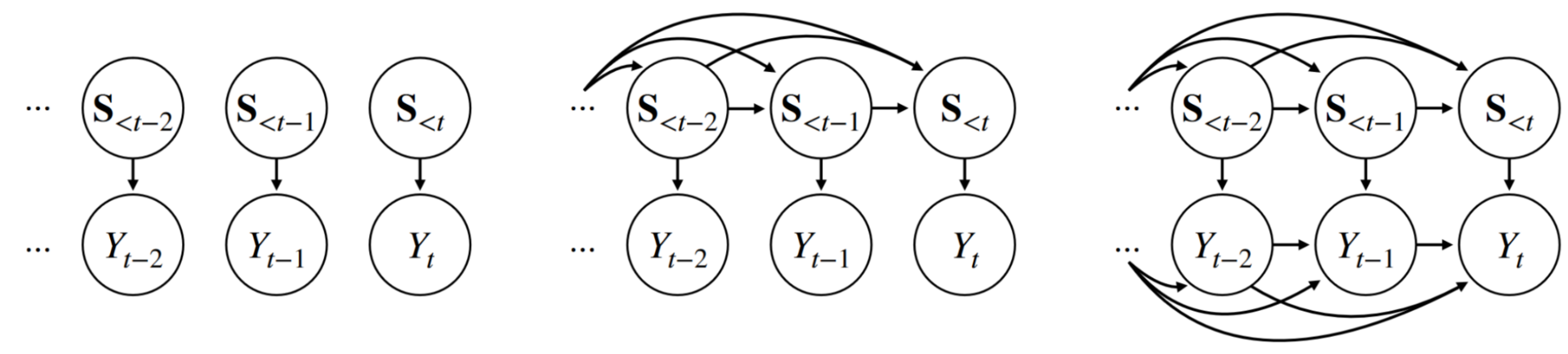
Analysis by basin ⇒ <u>Case study</u> of Hurricane Jose (2017)



- We interpret high LPD as a TC which is *"convectively primed" for RI*.
- Hurricane Jose was subject to high vertical wind shear (cause of RW) near Sept 9, which our model does not account for.

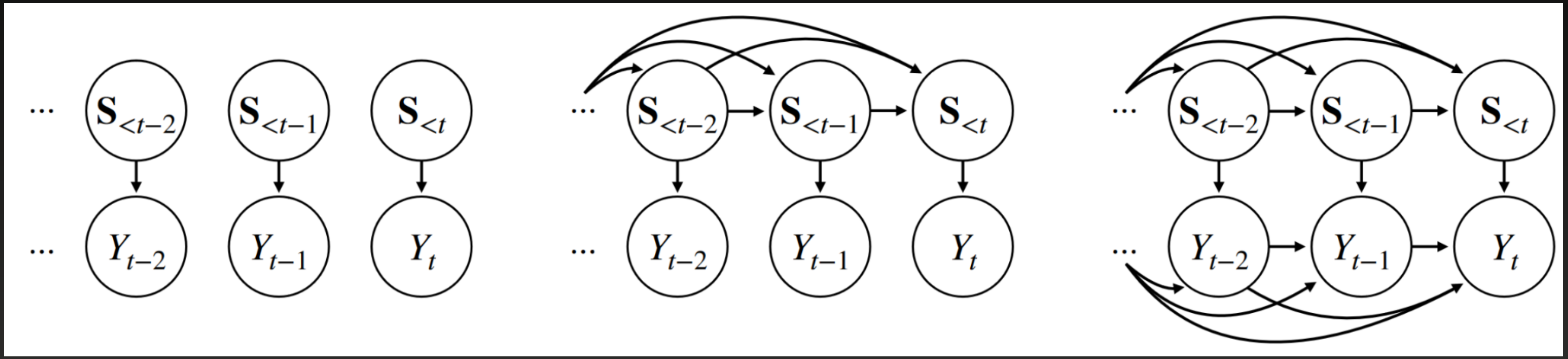
Summary: Detecting Distributional Differences

$$H_0 : p(\mathbf{s}_{<t}|Y_t = 1) = p(\mathbf{s}_{<t}|Y_t = 0)$$



- We have proposed a two-sample test for D.I.D sequence data $\{(\mathbf{S}_{<t}, Y_t)\}_{t\geq 0}$ with interpretable diagnostics. Two key ideas:
  - a test statistic based on the posterior difference p(Y=1|**s**)-p(**s**), estimated via a suitable regression method;
  - a bootstrap test where we estimate the marginal distribution of $\{Y_t\}_{t\geq 0}$; consistency guarantees asymptotic validity

Potential Extensions and Future Work



- Extend inputs S to include other functional features and data sources.
- Can extend to a *conditional* test H0: $p(s|Y=1,z) = P(z|Y=0,z)$ by considering the posterior differences $P(Y=1|s,z)-P(Y=1|z)$.

Acknowledgments

- Trey McNeely (CMU, now Microsoft Research)
- Galen Vincent (CMU, now Maxar)
- Dr Kimberly M Wood (MSU, Geosciences)
- Dr Rafael Izbicki (UFSCar)

This work is funded in part by NSF DMS-2053804 and NSF PHY-2020295.