The Extremal Dependence Index (EDI): More Than a Tool for Rare Event Forecast Evaluation?

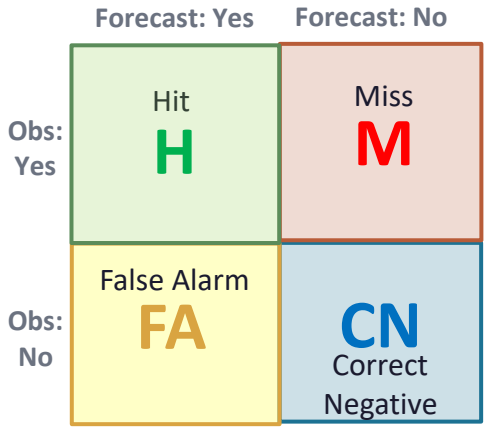
Testing the applicability

Iris Odak, Sara Anđela Perić, Jakov Lozuk

Croatian Meteorological and Hydrological Service

Mon 7 Sep 2026, 15:00 CEST | Room Quest

Motivation and theoretical background…



N=H+M+FA+CN
n (event)=H+M
p = n/N
empirical observed base rate

Probability of detection:
Fraction of observed events that were correctly forecast
POD = H/(H+M)

Probability of false detection:
Fraction of observed non-events that were incorrectly forecast as events
POFD = FA/(FA+CN)

Peirce skill score PSS:
(AKA the Hanssen–Kuipers discriminant or true skill statistic)
Detection achieved beyond false detection
PSS=POD−POFD

Extremal Dependence Index :
Log-scaled separation between detection and false detection, designed for rare events

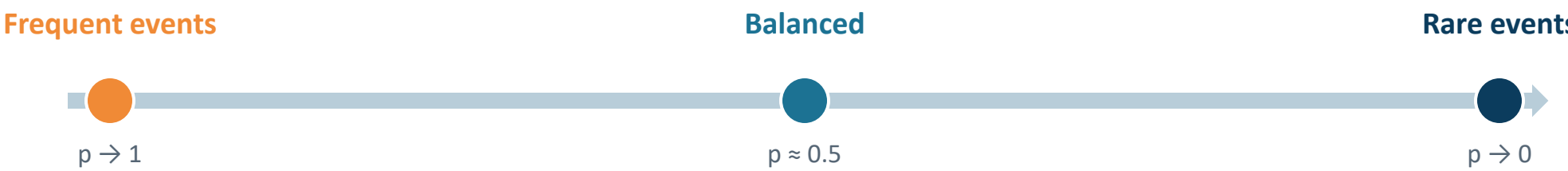
$$EDI = \frac{\ln POFD - \ln POD}{\ln POFD + \ln POD}$$

Symmetric Extremal Dependence Index:
Extends EDI by including both event and non-event rates -more symmetric assessment

$$SEDI = \frac{\ln POFD - \ln POD - \ln(1 - POFD) + \ln(1 - POD)}{\ln POFD + \ln POD + \ln(1 - POFD) + \ln(1 - POD)}$$

Event frequency changes how categorical scores behave

Operational verification must compare categories across the full spectrum—not only rare extremes.



Conventional scores may lose discrimination as the event approaches the rare-event limit.

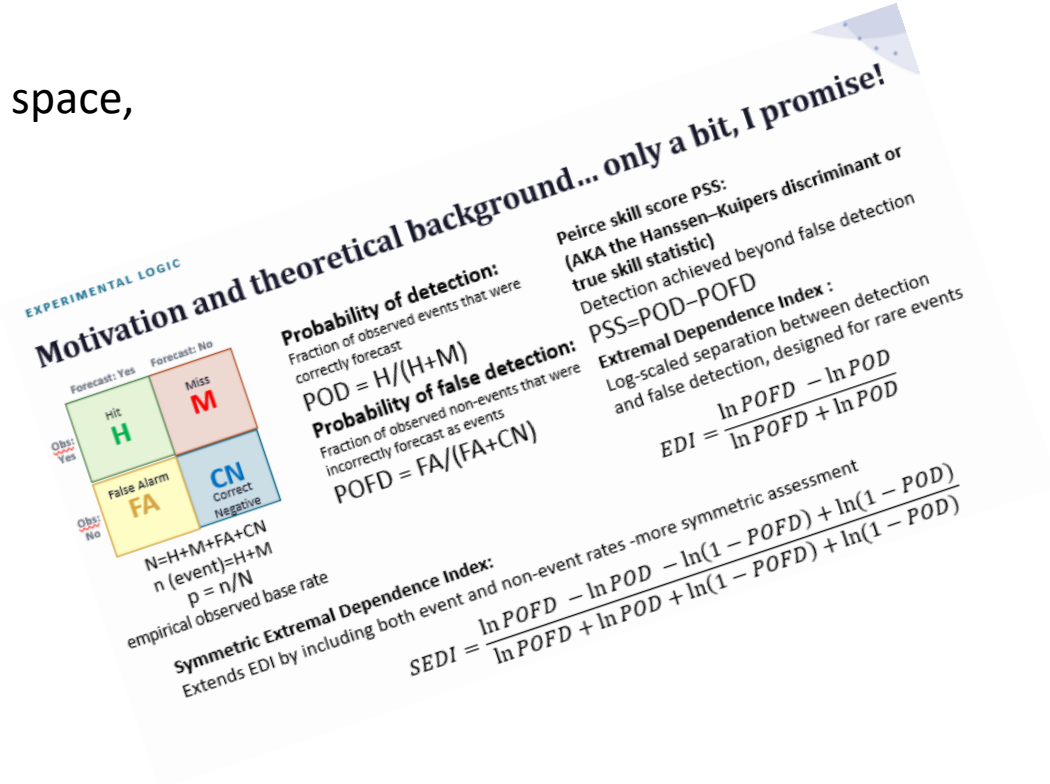
Rare-event measures avoid that limit—but their behaviour away from it still has to be tested.

Can one measure support a consistent comparison across the full frequency spectrum?

Score behaviour is predetermined—but not always intuitive

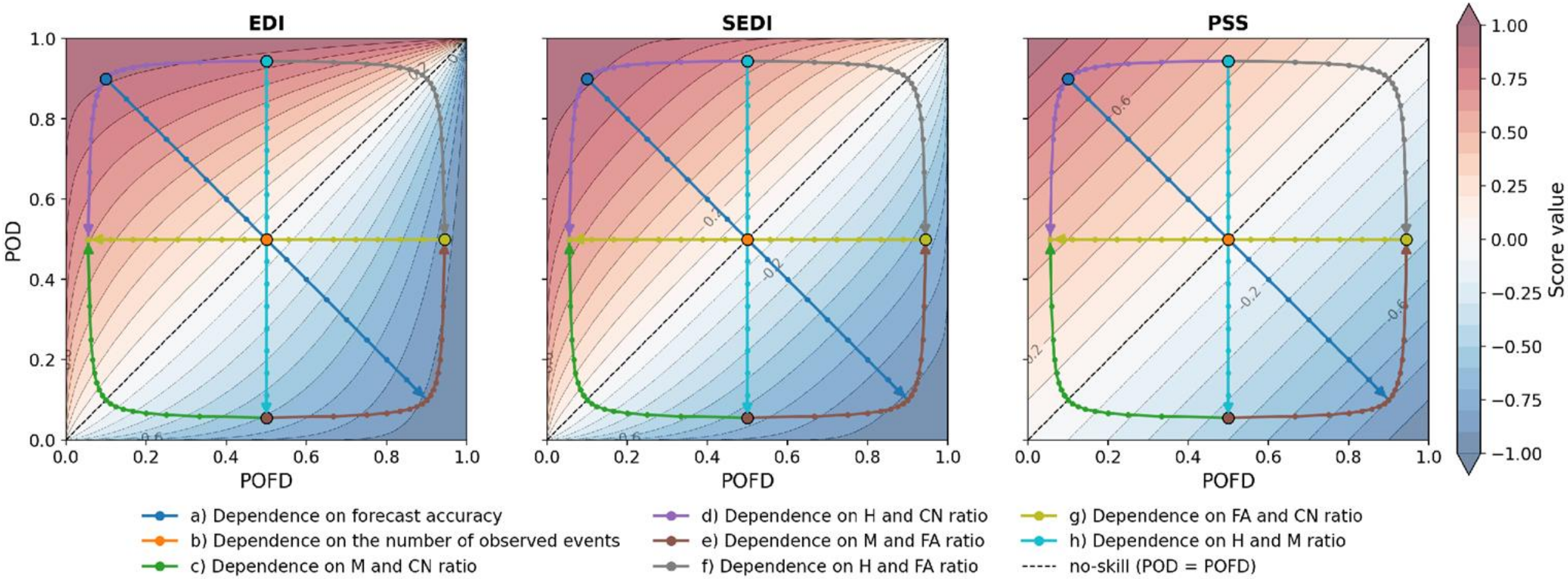
Operational verification must compare categories across the full spectrum—not only rare extremes.

- Controlled scenarios trace trajectories through contingency-table space, POD/POFD space.
- Each trajectory varies one structural feature: forecast accuracy, event frequency, or error balance.
- N is fixed by design - scenarios isolate score responses to changes in contingency-table structure rather than sample size
- EDI, SEDI and PSS are compared along every trajectory.
- TS, ETS and SEDS are retained as contextual comparators.

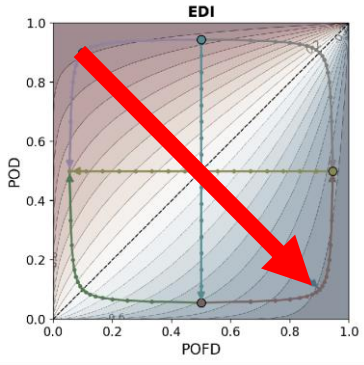


The scenarios as paths through POD-POFD space

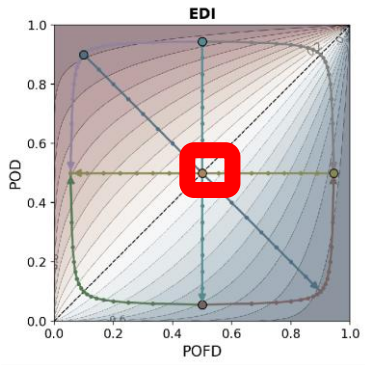
All measures share the no-skill line; their sensitivity away from it differs.



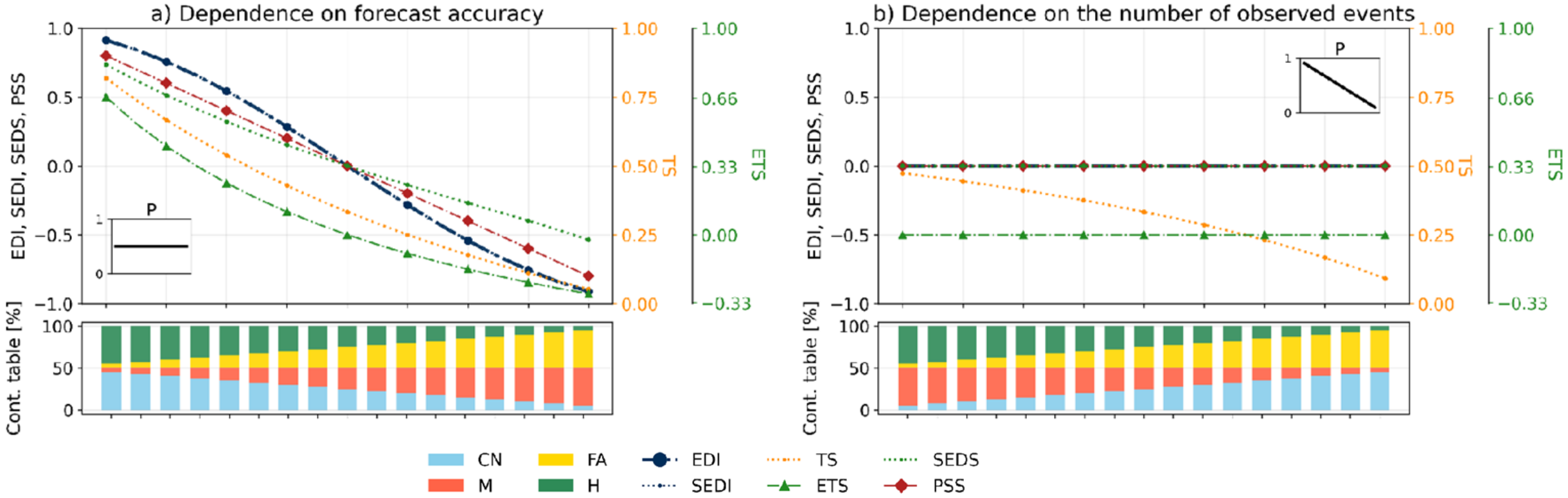
- EDI, SEDI and PSS are compared along every trajectory. .

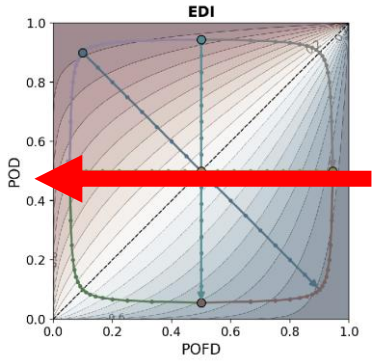


A score should react to skill not to event count alone



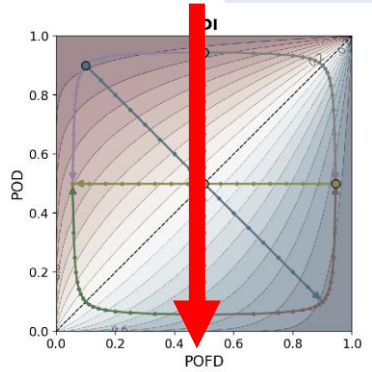
All three respond to forecast accuracy; No changes along the constructed no-skill path.

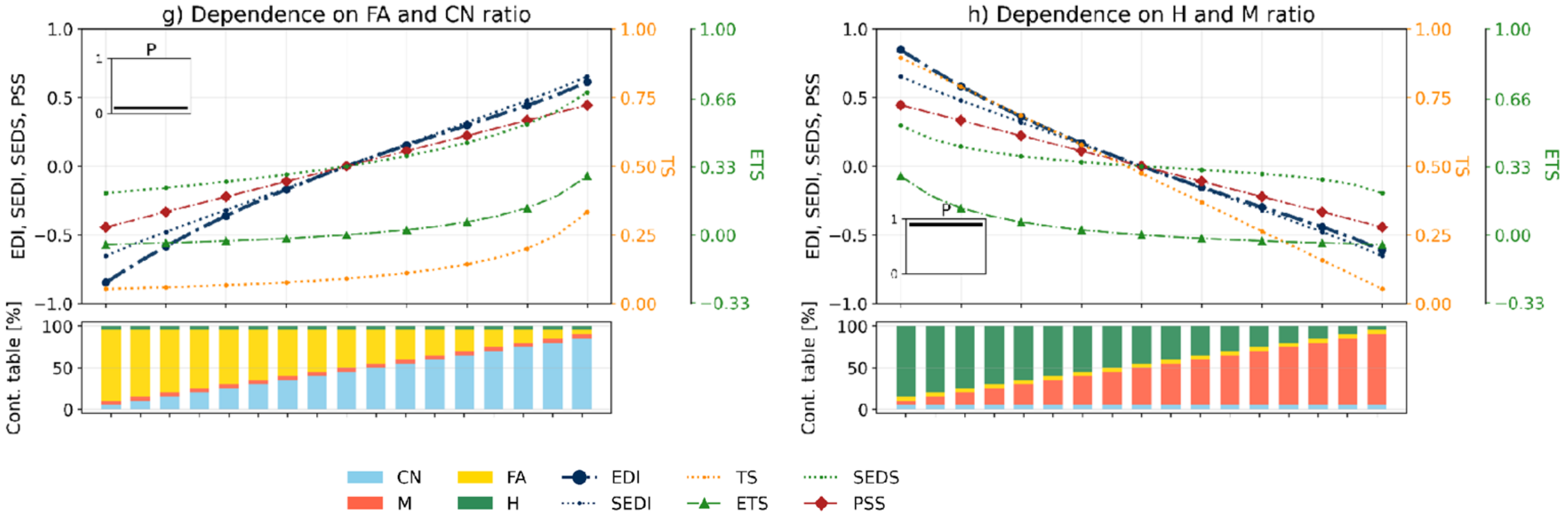




EDI and SEDI respond onto how errors are redistributed

The scores agree on direction, but their dynamic range differs.





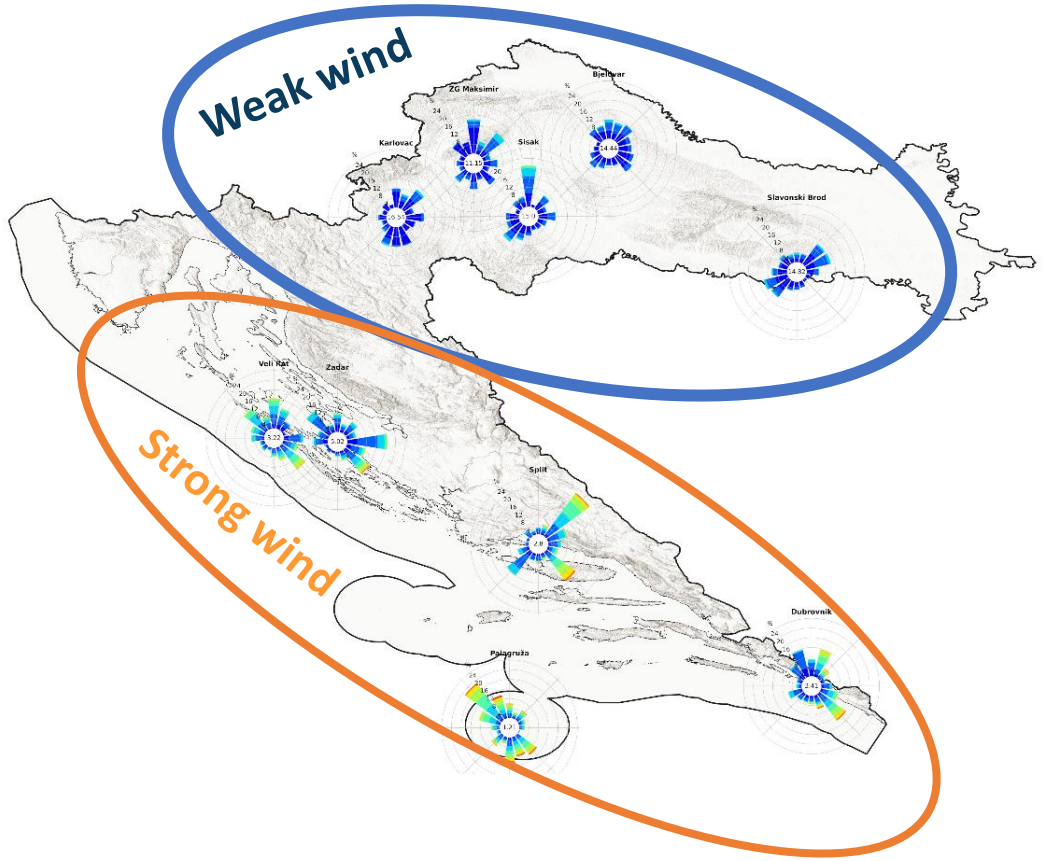
DATA

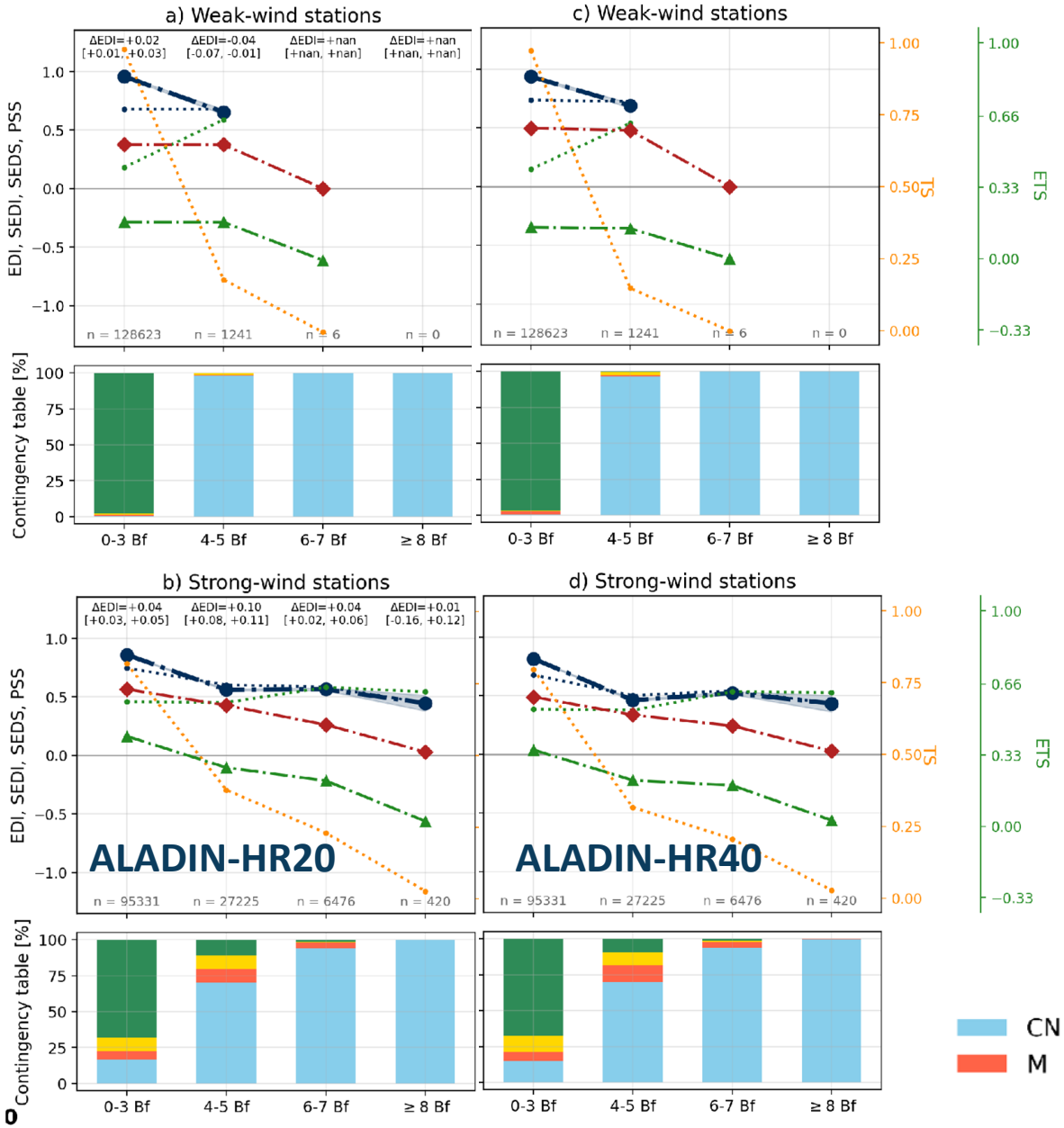
Operational ALADIN wind forecasts across Croatia

- Two ALADIN configurations: HR20 and HR40, differing in resolution and physics
- Both configurations: hourly output, 72-h lead time
- 10-minute mean wind speed observations, DHMZ stations, 2023
- Stations grouped into weak-wind and strong-wind regimes

Why two regimes?

The same wind category can be common in one climatology and rare in another.





- Verified against Beaufort-scale categories:
  0–3, 4–5, 6–7, ≥8 Bf
  = gentle , moderate, strong, gale force
- One-versus –rest category split

EDI and SEDI remain informative while event counts are adequate
Model differences are generally smaller than category and regime effects.

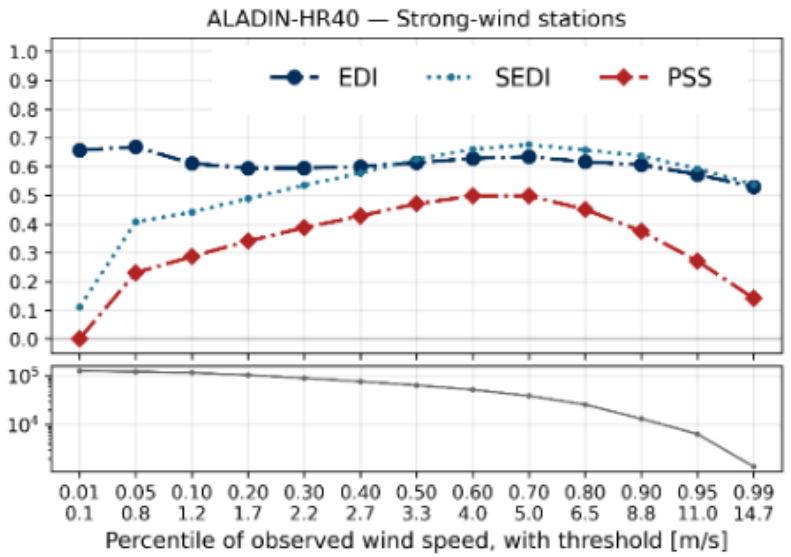
Applicability ≠ Estimability

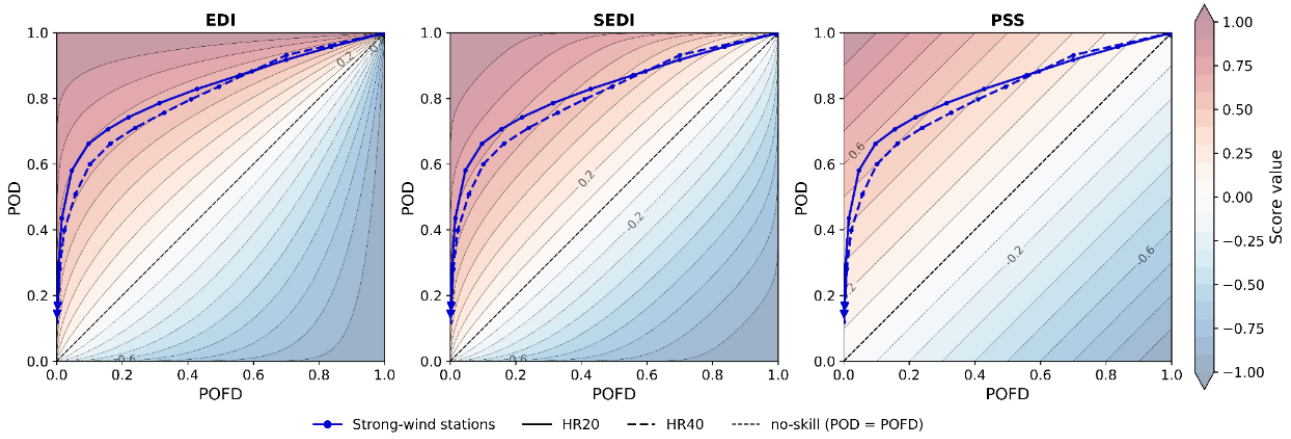
At weak-wind sites, the rarest bins disappear because the observed event count collapses—not because the score has a trivial rare-event limit.

EDI and SEDI remain non-trivial for the very common categories

Continuous exceedance-threshold analysis: strong wind stations

ALADIN-HR40





Diagnostic POD vs POFD

The effect of event vs non-event split?

Scores well defined

Common end:

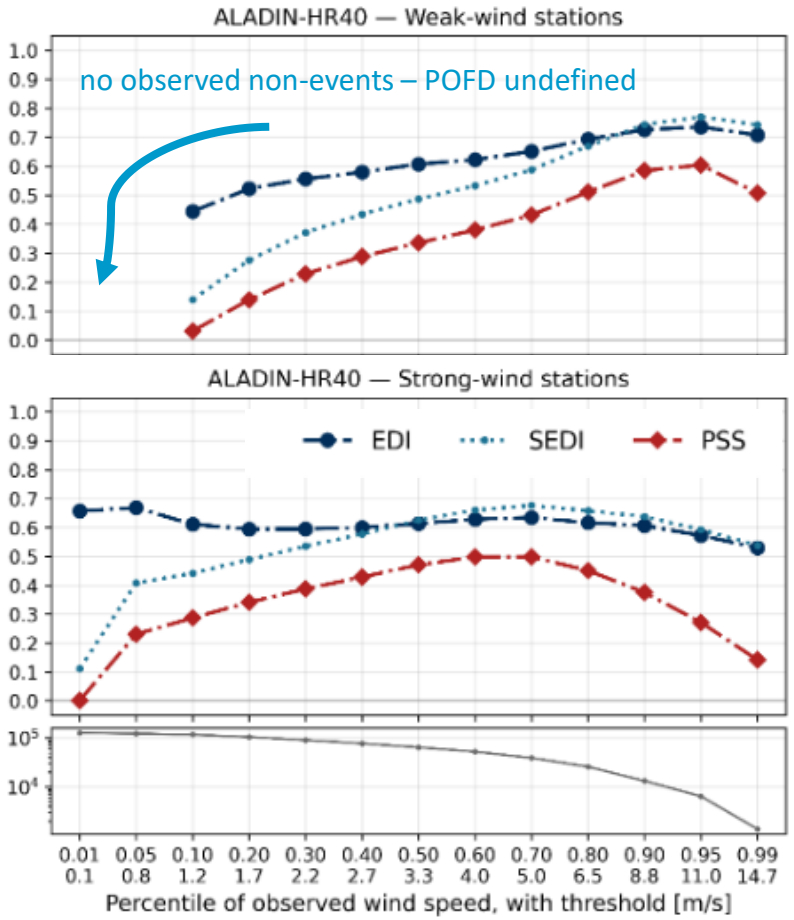
- POD,POFD near 1
- Forecast predominately Yes
- EDI and SEDI differ
- SEDI & PSS low for weak discrimination
- EDI remains high due to asymmetry

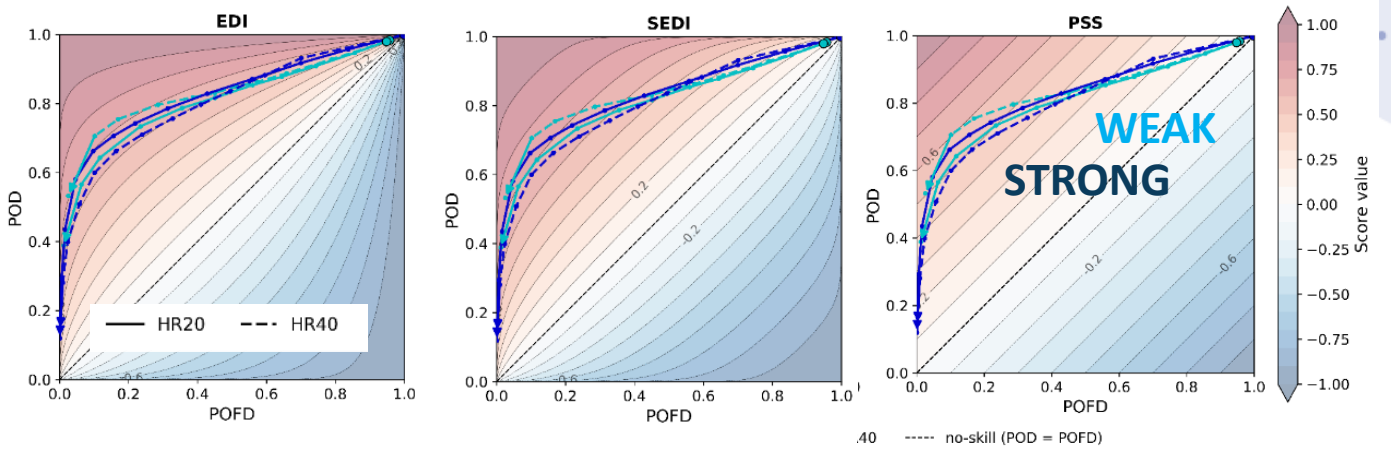
Rare end:

- POD drops when POFD is already low
- D loss outweights reduction in FD
- EDI&SEDI converge and decline
- Decline not due to degeneration of the score but because they asses POD-PODF trade-off

Continuous exceedance-threshold analysis – weak wind stations

ALADIN-HR40



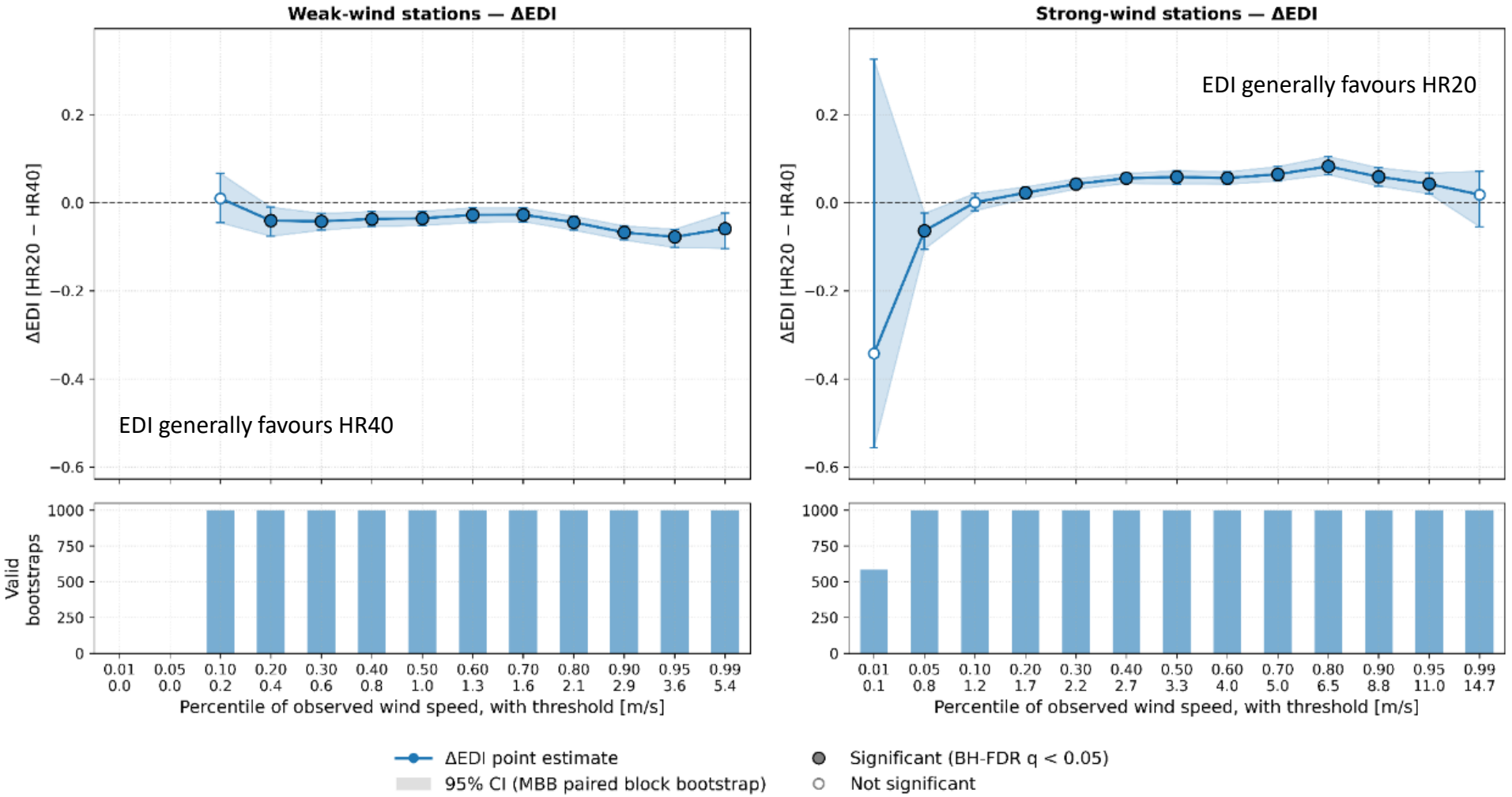


EDI and SEDI diverge toward common thresholds.

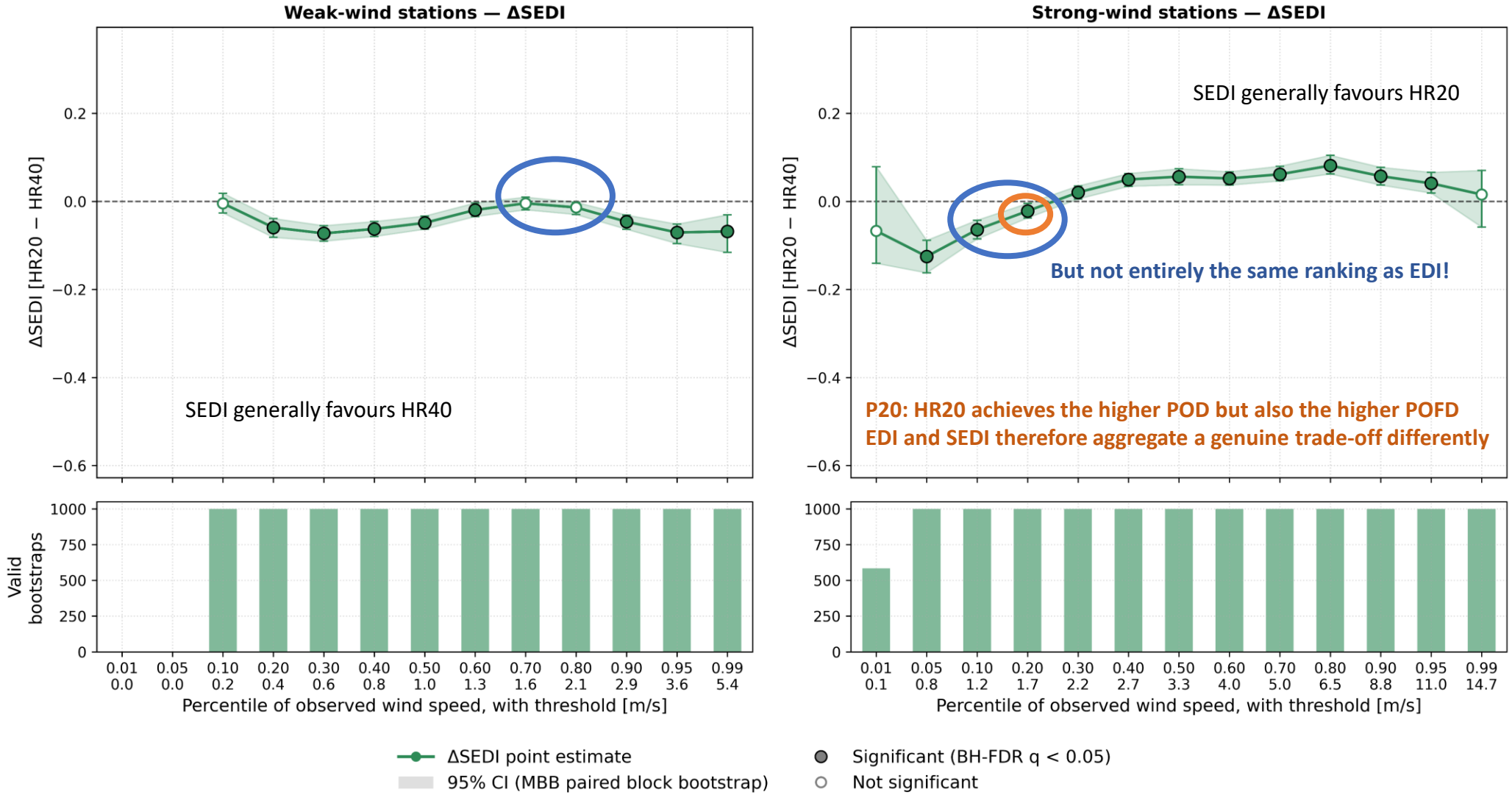
SEDI more closely follows the weak discrimination visible from the small separation between the two conditional rates

Undefined scores on the left side – limitations for all measures

Finally: which model configuration is favoured? Depends on wind regime and threshold!



Finally: which model configuration is favoured? **Depends on wind regime and threshold!**



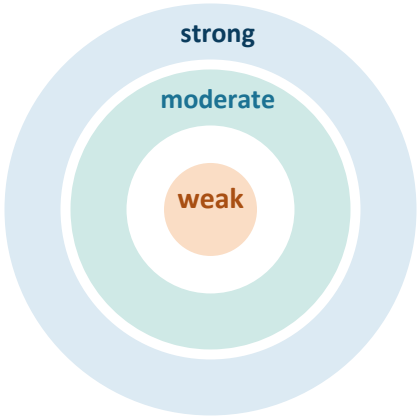
Category design determines the safest overall score

If frequency and purpose are known: common → PSS | transparent linear baseline | focal extreme → EDI | event orientation is intentional | simply rare → SEDI | no reason to privilege one label over its complement

Operational one-versus-rest verification of wind?

Speed only

3 classes; one may dominate / be rare

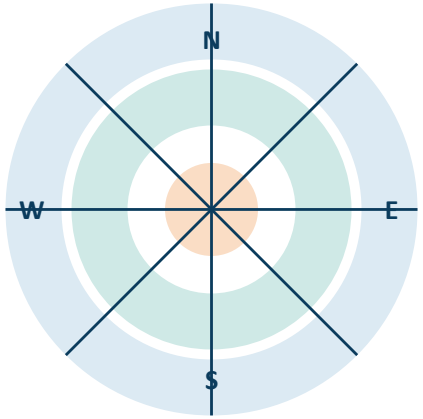


Safest default: SEDI

split by direction

Speed × direction

3 × 8 = 24 focal cells - minority events



EDI or SEDI

EDI fits when each cell is the focal event

For every score, report p and N: frequency gives context, sample size gives support (event count ≈ pN).

TAKE-HOME MESSAGE

EDI is useful beyond extremes—but context still decides

01 Both EDI and SEDI remain informative beyond rare events

Across theoretical paths and real thresholds, it stays useful over a broader frequency range than its original niche suggests.

02 No single score is context-free

PSS suits common events; EDI a focal extreme or minority cell; SEDI is safer when frequencies are unknown.

03 Never show the score alone

Report p and N, and use POD and POFD to explain why measures or model rankings diverge.

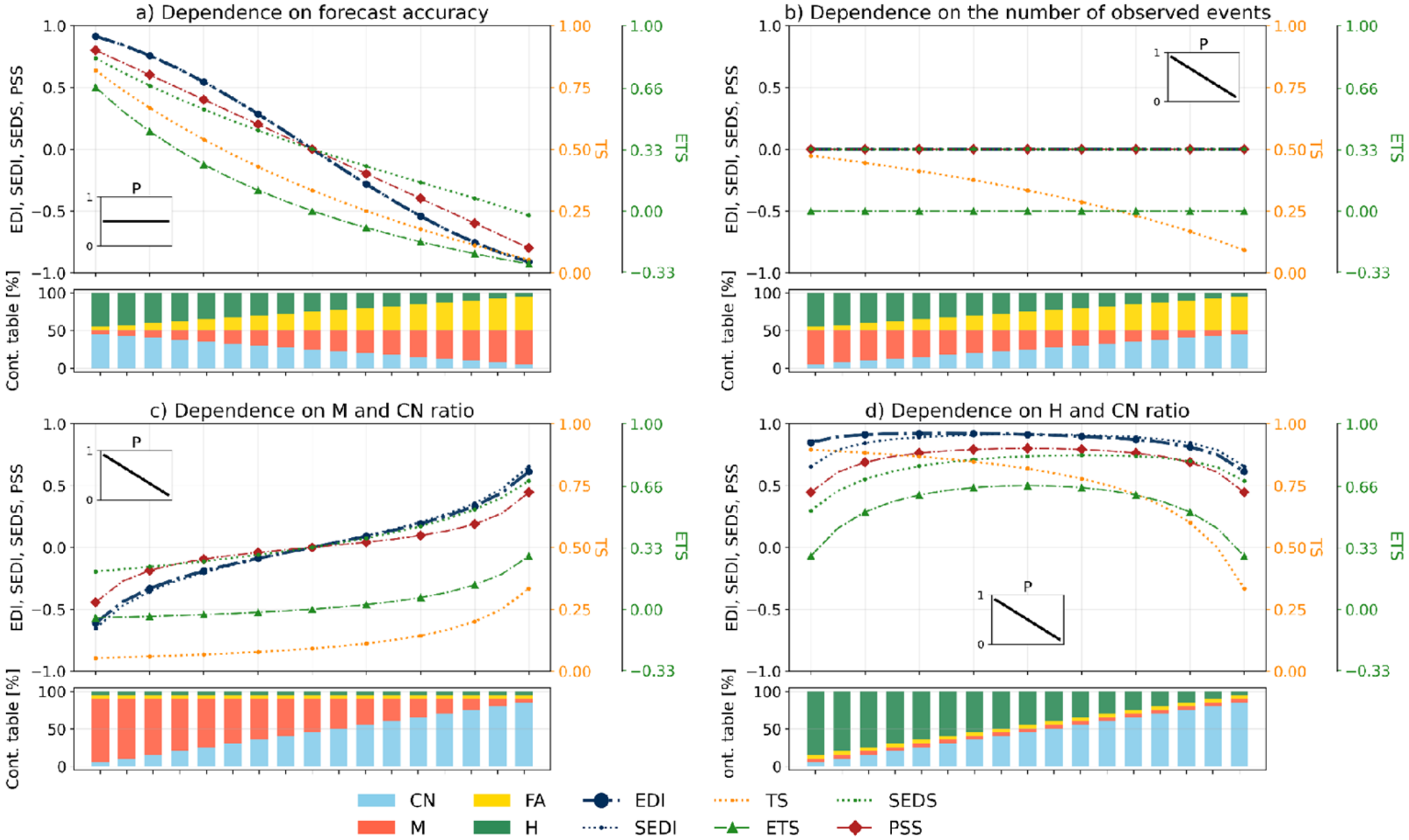
Final formulation: EDI/SEDI can be more than a tool for rare-event forecast evaluation, but it is not a stand-alone verdict

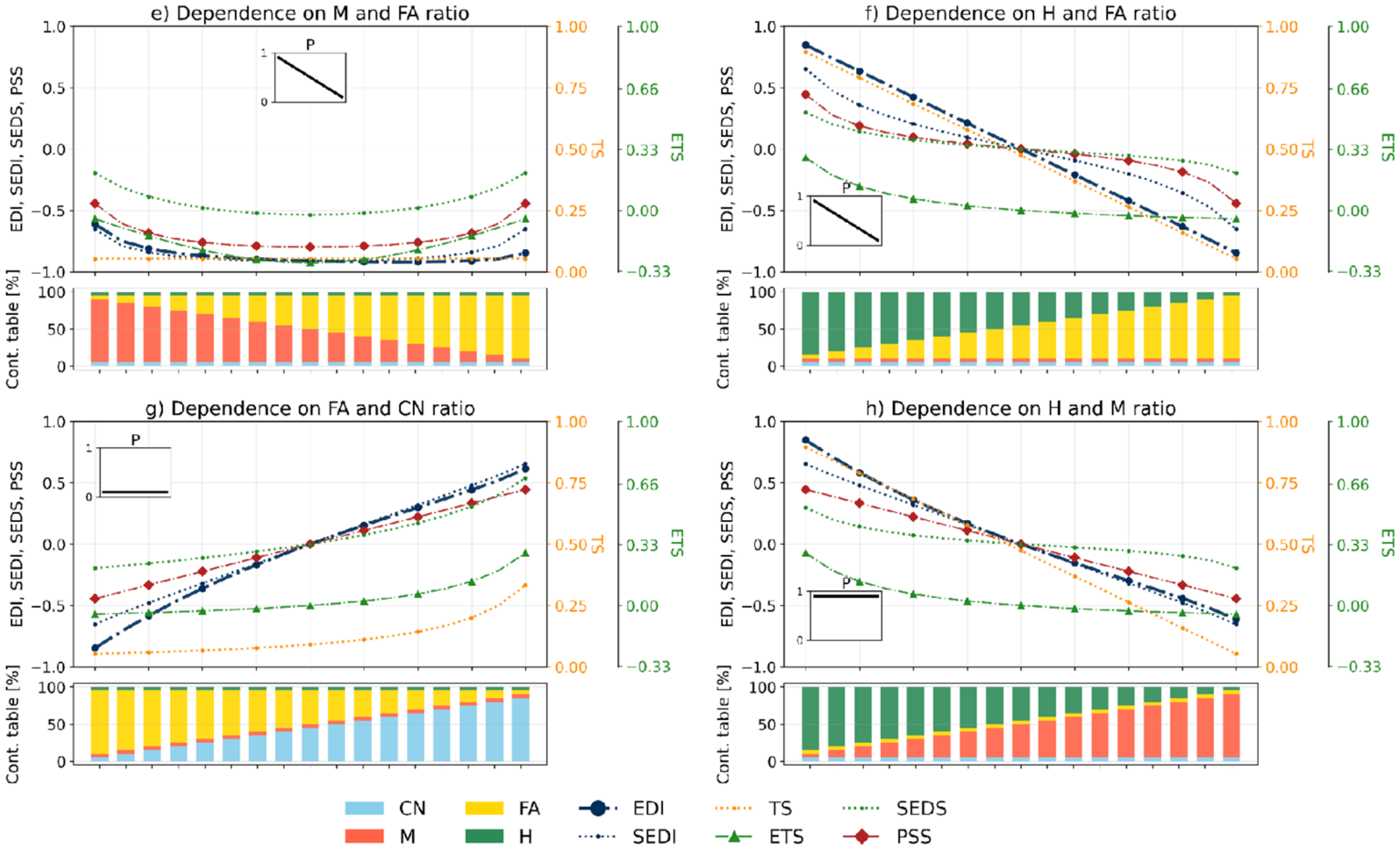
What do you think?

questions, comments and opinions welcomed...

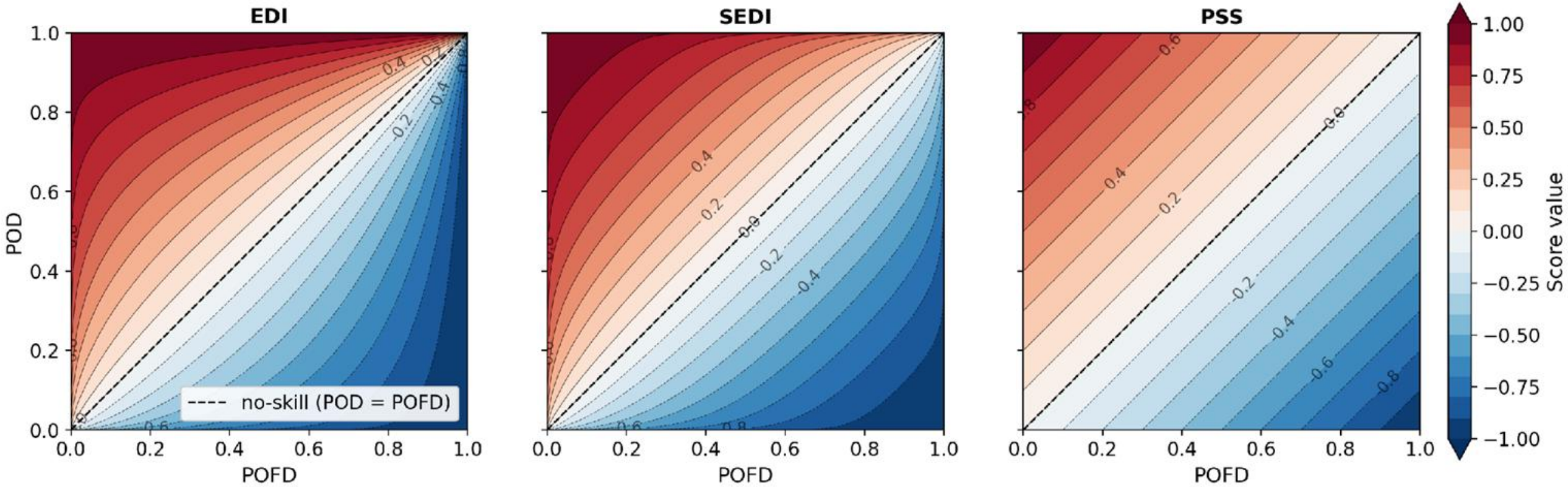
odak@dhz.hr

Thank you!


a) Dependence on forecast accuracy
b) Dependence on the number of observed events
c) Dependence on M and CN ratio
d) Dependence on H and CN ratio
EDI, SEDI, SEDS, PSS
TS
ETS
Cont. table [%]
P
CN
M
FA
H
EDI
SEDI
TS
ETS
SEDS
PSS


e) Dependence on M and FA ratio
f) Dependence on H and FA ratio
g) Dependence on FA and CN ratio
h) Dependence on H and M ratio
EDI, SEDI, SEDS, PSS
TS
ETS
Cont. table [%]
P
CN
M
FA
H
EDI
SEDI
TS
ETS
SEDS
PSS

EDI, SEDI and PSS are functions of POD and POFD only



Paired Moving Block Bootstrap (MBB)

- The paired MBB is used to estimate the **95% confidence intervals** of the differences between HR20 and HR40.
- The differences are defined as **ΔEDI = EDI(HR20) − EDI(HR40)** and **ΔSEDI = SEDI(HR20) − SEDI(HR40)**.
- Seven-day moving blocks are used to preserve the **temporal dependence** in the data.
- The blocks may overlap and are sampled **with replacement** until the original sample length is reconstructed.
- The **same resampled dates** are used for HR20 and HR40, preserving their paired structure.
- EDI and SEDI are recalculated for each of **1000 bootstrap samples**.
- The 95% confidence interval is obtained from the **2.5th and 97.5th percentiles** of the bootstrap distribution of ΔEDI or ΔSEDI.

Observed HR20 + HR40 time series
↓
7-day moving blocks
overlapping + sampled with replacement
↓
Paired bootstrap samples
same dates for HR20 & HR40
↓
Calculate ΔEDI / ΔSEDI
↓
1000 bootstrap replicates
↓
95% CI
2.5th–97.5th percentile

Key idea: The paired MBB preserves both **temporal dependence** and the **HR20–HR40 pairing**.

Paired Block Permutation Test and FDR Correction

- The paired block permutation test is used to determine whether the observed **difference between HR20 and HR40 is statistically significant**.
- The null hypothesis is that **ΔEDI = 0** or **ΔSEDI = 0**, meaning that the two model configurations do not differ systematically.
- The common dates are divided into **fixed, non-overlapping seven-day blocks**.
- In each permutation, every block is either **kept unchanged or swapped between HR20 and HR40**.
- The same swap pattern is applied to EDI, SEDI, and all **13 wind-speed percentiles**, preserving their paired structure.
- A total of **5000 permutations** are used to construct the null distribution.
- The two-sided p-value is calculated from the proportion of permutations with a difference at least as extreme as the observed difference.
- **Benjamini–Hochberg FDR correction** is then applied to the 13 percentile-based tests, separately for each wind-speed group and each metric, using **α = 0.05**.