Response to Reviewers

We thank both referees for their careful and constructive reviews. Before answering each comment, we summarize four changes that affect several responses. Throughout the responses, numbers refer to the revised catalogue and to the two-regime solution now selected, and figure numbers refer to the revised manuscript.

1. **Revised event catalogue.** Prompted by RC1's comment on the definition of independent events, we audited the catalogue for duplicate reports with an uncertainty-aware search followed by source-by-source verification (new **Fig. S1**). Fifty-three reports were confirmed as duplicates of other events and removed, reducing the catalogue from 1185 to 1132 events (540 tornadoes, 592 waterspouts). All analyses were recomputed with the revised catalogue.
2. **Corrections to the analysis code.** During the revision we checked the code step by step, which revealed four implementation issues that have now been corrected. (i) A column-screening step had excluded 11 of the 54 parameters of **Table S1** from the event-level analyses of Sect. 3.2; all analyses now use the same 432 predictors (54 parameters × 8 descriptors). (ii) In the evaluation of the candidate regionalizations, metrics were computed within each cross-validation fold using only the labels present in that fold and then averaged, which can overestimate the scores when a spatial block contains a single regime; they are now computed from the pooled out-of-fold predictions, as in Sect. 3.2. (iii) In the hazard calculation, 48 tornadoes reported at the coast, just outside the clipped domain, had not been assigned to any regime; they are now assigned to the nearest cell. (iv) In the hexagon-level maps, the empirical-Bayes prior had the same weight in every cell, which produced spurious extremes in small clipped coastal cells; the prior is now expressed for a full cell (Sect. 2.6). Corrections (i) and (ii) change the reported scores but not the regionalizations; corrections (iii) and (iv) change the hazard estimates and maps but not the ranking of the regimes.
3. **Selected regionalization.** The criteria used to select the final partition are unchanged (Sect. 2.5.3). Applied to the revised catalogue and the corrected scores, they now favour the two-regime partition. At 50 km, k = 2 obtained the highest composite score (0.75, compared with 0.57–0.64 for k = 3–6) and the highest dummy-relative skill averaged over the full and tornado-only samples (0.53, compared with 0.40–0.50), and it ranked first at all three cell sizes. For the full sample alone, skill varied little with k (0.44–0.51); the choice therefore rests mainly on the tornado-only subset used in the hazard analysis, where k = 2 is clearly best (0.60, compared with at most 0.49), and on parsimony. Because Ward clustering is hierarchical, the partitions are nested: the two-regime solution keeps the western–southwestern regime (regime 1) unchanged and merges the northern–eastern regime and the central–southeastern belt (regimes 0 and 2 of the discussion version) into a single regime 0. The new solution is therefore a coarser level of the same regional structure. The three-regime partition is shown in the Supplement for comparison (**Fig. S8**), and where the referees refer to specific regimes of the discussion version, our responses indicate the corresponding current regime.
4. **Reorganized Methods.** Following RC1 (major comment 4) and RC2 (major comment 1), Sects. 2.4 and 2.5 have been rewritten around the questions each step addresses, with a workflow figure (new **Fig. 1**), and the full settings have been moved to the Supplement.

Line numbers refer to the clean version of the revised manuscript.

Reviewer 1:

General assessment

The manuscript presents an interesting and potentially valuable contribution to the understanding of tornado and waterspout climatology and wind-speed hazard over peninsular Spain and the Balearic Islands. The updated event database, the environmental analysis based on ERA5, the objective spatial regionalization and the subsequent exceedance analysis provide a broad and ambitious framework. I also find the manuscript well contextualized, with a large number of relevant references, and the general presentation of the results is quite easy to follow.

In particular, I find the results section substantially clearer than the methodology section. The regional patterns and the distinction between tornado occurrence and high-threshold wind-speed exceedance are interesting and potentially useful. The paper therefore has good potential.

My main concern is not the general scientific interest of the work, but rather the complexity and presentation of the methodology, and, consequently, the difficulty for the reader to follow exactly how the final results are obtained. Several methodological decisions are introduced in a rather compressed way, while their practical meaning only becomes clear much later in the Results section. In my opinion, the manuscript would benefit considerably from a clearer and more pedagogical presentation of the methodological workflow, together with some additional discussion connecting the statistical results with the underlying meteorology.

I would therefore recommend major revision. Most of the issues below appear addressable without changing the general framework of the study.

Response:

We thank the referee for the positive assessment and for the detailed and constructive comments. We agree that the methodology was presented in a compressed way. The Methods have been reorganized around the questions each step addresses, with a workflow figure and explicit links to the results (see our responses to major comments 3–5), and the main changes affecting several responses are summarized in the General Response above.

Major comments

1. The methodological framework should be introduced and explained more clearly

The last paragraph of the Introduction introduces the footprint-based framework of Balbi and Barbieri (2018), but the reader is only given a relatively general description of it. Since this framework is an important part of the transition from climatology to hazard in the present study, I think it would be useful to explain already in the Introduction, in somewhat more concrete terms, which elements of the original framework are adopted and which are adapted for the Spanish case.

For example, the authors could briefly explain how occurrence rate, tornado intensity, path geometry and the distribution of wind speeds are combined to obtain the exceedance probability. This would help the reader understand from the beginning what the final objective of the study is and how the different methodological parts are connected.

Response:

We agree. The penultimate paragraph of the Introduction ***(L63-75)*** now explains the framework in concrete terms. It describes the four elements that are combined to obtain the exceedance probability: the areal occurrence rate of each intensity class, the expected footprint area derived from the path-length and width distributions, the fraction of the footprint exceeding each wind-speed threshold, and the Poisson model that converts the annual exceedance rate into EP(υ, T). It then states which elements are retained and which inputs are adapted to the Spanish case: regime-based occurrence rates derived from ERA5, path distributions fitted to the local catalogue with Brooks (2004) scaling for EF3, and hexagonal mapping with empirical-Bayes smoothing. The new workflow figure (**Fig. 1**) places this step within the complete analysis.

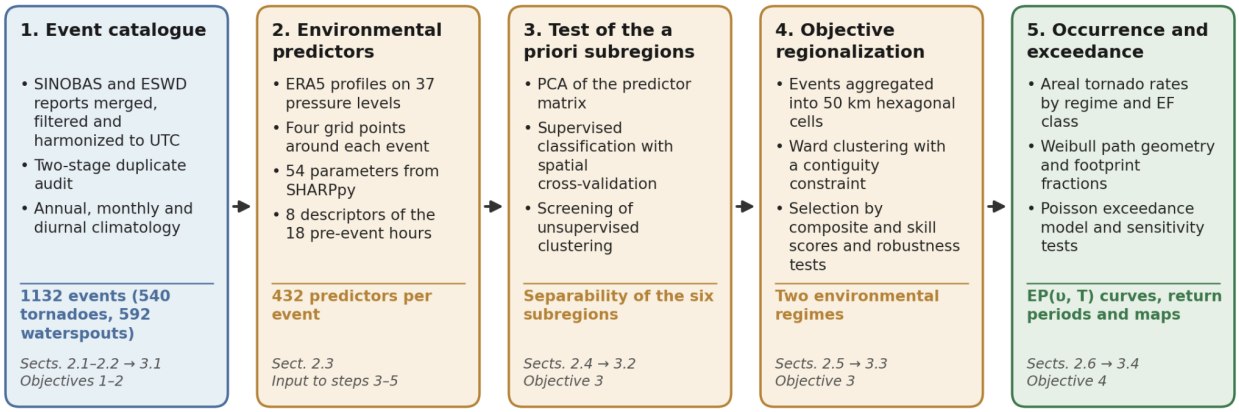


Figure 1. Workflow of the study. Each box lists the main operations of one step, its output (bold) and the Methods and Results sections and objectives to which it corresponds. Step 1 (blue) builds the event catalogue; steps 2-4 (tan) characterize the pre-event environments, test the a priori subregions and derive the objective regionalization; step 5 (green) translates the regime-dependent tornado climatology into wind-speed exceedance.

2. The definition of an independent event deserves more attention

The treatment of duplicate events should be clarified. The manuscript states that reports within 50 km and 2 h are considered duplicate candidates, but it is not sufficiently clear how the final decision is made in such cases.

This point is particularly important for tornado families, outbreaks, or sequences of tornadoes produced by the same convective system. Two tornadoes occurring within 50 km and 2 h are not necessarily duplicate observations of the same tornado. They may represent two physically different tornadoes associated with the same storm or convective system.

Related to this issue, Section 2.6 describes the occurrence process as involving "rare and approximately independent events". The authors should discuss whether this assumption is fully compatible with possible tornado families or outbreaks containing several tornadoes. If some degree of dependence between events is expected, the implications for the occurrence and exceedance calculations should at least be discussed.

Response:

We thank the referee for raising this important point, which led us to audit the catalogue in detail.

(i) Duplicate criterion: duplicate removal now follows a two-stage procedure, described in Sect. 2.1 ***(L117-131)*** and in the new **Fig. S1**. The first stage applies the 50 km/2 h criterion of Rodríguez and Bech (2021) only to consecutive reports of the same intensity class, together with a one-by-one screening of SINOBAS reports and a check of the merged pairs against the original reports. The second stage, added in this revision, never merges candidates automatically. Each report's declared time precision was converted into an explicit uncertainty; candidate pairs were searched within 50 km and 2 h plus their combined uncertainty, first within and then across phenomenon types and intensity classes; and each candidate group was verified against the mapped coordinates, the original source reports and independent contemporary documentation. Reports were kept as distinct events by default and merged only when an external source confirmed that they described the same vortex, so that tornadoes belonging to the same family or outbreak are retained as separate events (for example, the Barcelona outbreak of 7 September 2005 and the Valencia outbreak of 29 October 2024). This second stage removed 53 further duplicates (General Response, point 1); widening its thresholds by 50 % added only 22 (distance) and 6 (time) candidate pairs, none involving an EF2 or EF3 event.

(ii) Independence: we agree that tornadoes from the same storm or outbreak are not strictly independent, so annual counts can be more variable than a Poisson process implies. This affects the uncertainty rather than the central estimate: the exceedance probability depends on the mean annual rate, and for the small values obtained here ($\lambda T \ll 1$) $EP \approx \lambda T$ regardless of clustering. For a given mean rate, clustering would in fact slightly reduce the probability of

at least one exceedance, so the Poisson assumption is conservative. Consistently, the cell-level tornado counts are overdispersed by a factor of about 5.3–8.5 with respect to a Poisson model (Sect. 3.4), and the year-block bootstrap, which resamples whole years, retains within-year clustering in the uncertainty ranges. This is now discussed in Sect. 2.6 ***(L411-418)***.

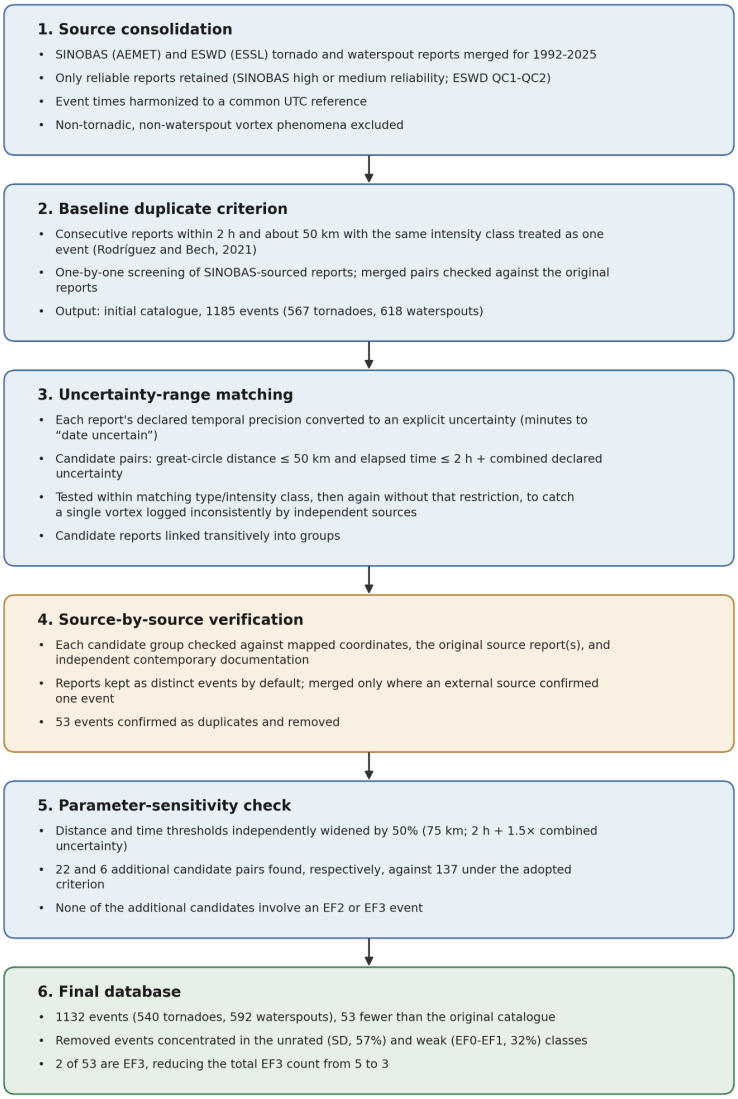


Figure S1. Construction and duplicate audit of the tornado and waterspout catalogue. Steps 1-2 (blue) reproduce the merge and baseline duplicate screening described in Sect. 2.1, yielding the initial 1185-event catalogue. Steps 3 and 5 (blue) extend this to an uncertainty-aware search across phenomenon type and Fujita class, with a check of the search thresholds' sensitivity; step 4 (tan) is the accompanying manual, source-by-source verification; step 6 (green) is the resulting corrected catalogue used throughout the study.

3. The terminology "regime", "region", and "class" should be defined early and used consistently

The manuscript uses the terms region, subregion, regime, and class, but their precise meaning is not always immediately clear to the reader. In particular, the use of regime sometimes feels somewhat abstract and the reader has to discover its meaning progressively.

I suggest defining these terms explicitly before they are used extensively. For example, the manuscript could clearly distinguish between: the initial geographically defined regions; the objectively derived environmental/spatial regimes; and the intensity classes such as EF0, EF1, etc.

I also think it would be useful to explain why regime is preferred to region for the final spatial units, since the latter may sometimes be more intuitive.

Response:

We agree. A short paragraph at the start of Sect. 2 ***(L96-100)*** now defines the three terms before they are used: subregion for the six geographically defined a priori units (Sect. 2.2), regime for the spatial units derived objectively from ERA5 environments (Sect. 2.5), and class for intensity categories only. We prefer regime to region because the final units are defined by shared pre-event environmental conditions rather than by geographical criteria; region would suggest a purely geographical subdivision, even though each regime is mapped as a contiguous area. The terms are now used consistently throughout, and in the classification analyses the predicted categories are referred to as labels, so that class is reserved for intensity.

4. The PCA and clustering methodology needs considerably more methodological detail

This is probably the main issue I have with the manuscript.

Section 2.4 introduces a relatively complex sequence involving standardization, PCA, cumulative explained variance, DBSCAN, grid-based density analysis, Ward clustering, supervised classification, several performance metrics, normalization of MCC, a composite score and a dummy-relative skill score. Section 2.5 then introduces another sequence involving hexagonal aggregation, medians, imputation, standardization, PCA, Ward clustering with spatial connectivity, different grid sizes, different numbers of clusters, spatial cross-validation and several criteria for the final selection.

Although all these elements may be reasonable individually, their presentation is very compressed. It is difficult to reconstruct the complete workflow and understand which methodological choices are actually used to obtain the final three-regime solution.

The paragraph beginning around line 156 also needs more explanation. It would help to give a simple example showing what is meant by noise fraction, dominant-cluster fraction and the screening of candidate solutions.

More generally, I suggest that the authors consider reorganizing Sections 2.4 and 2.5 so that the methodological steps are presented in a more intuitive sequence and are directly connected to the corresponding results. At present, the reader is sometimes asked to remember a complex methodological procedure and only much later discovers why a particular step was important.

Response:

We thank the referee for this comment, which we agree identified the main weakness of the manuscript. Numbers below refer to the revised catalogue and to the two-regime solution now selected (General Response, points 1–3). We have restructured the Methods as follows:

(1) A new introductory paragraph and workflow figure (Sect. 2 and **Fig. 1**, ***L83-90***) summarize the complete procedure, linking each step to the objectives and to the corresponding results section.

(2) Section 2.4 now opens with the question it addresses (are the six a priori subregions environmentally distinct?) and is divided into four parts presented in the same order as the results in Sect. 3.2: dimension reduction (2.4.1), supervised test (2.4.2), scoring (2.4.3) and clustering screening (2.4.4). Each part explains why a step is needed before describing how it is done, for example why PCA is applied, why the cross-validation is spatially grouped, why several metrics are combined, and what the dummy reference and the skill score represent.

(3) Section 2.5 opens with the complementary question (which regional structure do the environments themselves support?) and is divided into aggregation into cells (2.5.1), constrained Ward clustering (2.5.2), evaluation and selection (2.5.3), extension to cells without events (2.5.4) and physical characterization of the regimes (2.5.5).

(4) The choices used for the final solution are now stated explicitly at the end of Sect. 2.5.3: Ward clustering of 50 km cells, 15 leading PCs (77.9 % of the variance), edge-contiguity constraint and k = 2 regimes.

(5) The paragraph formerly around line 156 now includes a numerical example of the noise fraction, the dominant-cluster fraction and the screening thresholds (Sect. 2.4.4; see also our response to comment 16).

(6) To keep the main text readable without losing detail, the full settings of both stages are listed in a new **Table S2**, and the metric formulas in Sect. S1 of the Supplement.

5. The spatial-connectivity constraint in the Ward clustering should be explained

The manuscript states that a spatial-connectivity constraint was used to prevent geographically fragmented regions. However, the exact way in which this constraint is imposed is not sufficiently clear.

This is an important methodological point because spatial coherence is one of the main arguments for selecting the final regionalization. The authors should explain more explicitly how neighbouring hexagons are defined, how the connectivity matrix/constraint is constructed, and how this constraint modifies the Ward clustering algorithm.

A simple schematic example would be very helpful.

Response:

We agree that this point needed a clearer explanation. Section 2.5.2 ***(L299-309)*** now explains the constraint in three steps.

(i) Neighbours: two occupied hexagonal cells are neighbours when they share an edge.

(ii) Constraint: all neighbouring pairs are stored in a binary, symmetric connectivity matrix (1 for neighbours, 0 otherwise).

(iii) Effect on Ward: the algorithm starts with each cell as its own group and at each step merges the two groups whose union least increases the within-group variance; with the constraint, only adjacent groups (those containing at least one pair of neighbouring cells) can be merged, so every regime is a contiguous set of cells (Guo, 2008). A schematic example with seven cells and the corresponding matrix has been added (**Fig. S4**). We also report that at 50 km all occupied cells form a single connected network, so the constraint is purely geographical, and that at 35 km one isolated cell is linked by the scikit-learn implementation through the environmentally closest cell.

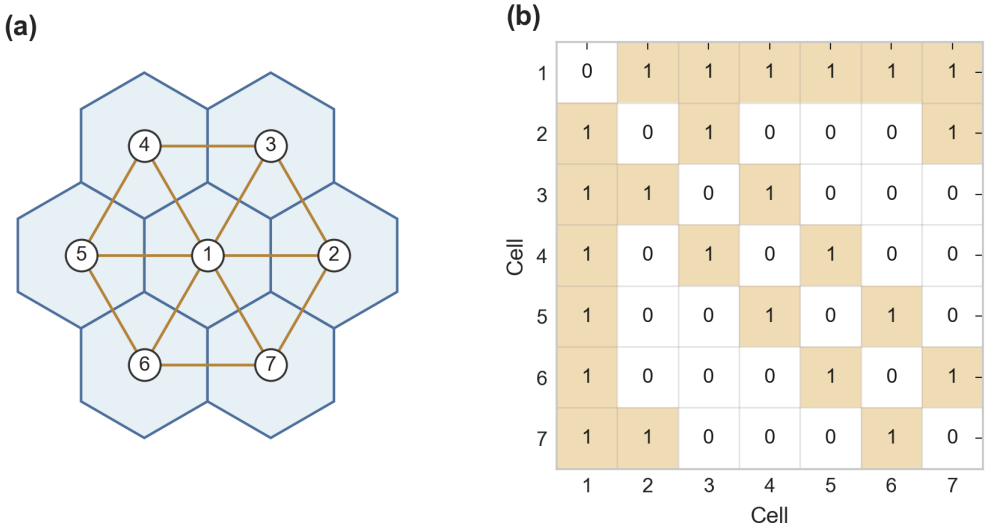


Figure S4. Edge-contiguity constraint used in the spatially constrained Ward clustering (Sect. 2.5.2), illustrated with seven hexagonal cells. (a) Two cells are neighbours when they share an edge (lines between cell centres). (b) Corresponding binary connectivity matrix (1, neighbours; 0, otherwise). With the constraint, Ward clustering can only merge groups that contain at least one pair of neighbouring cells. Cells 2 and 5, for example, are not neighbours and can belong to the same group only if it also contains cells that connect them, so every regime is a contiguous set of cells.

6. The relationship between the methodological complexity and the physical interpretation should be strengthened

The final regionalization is interesting, but the physical interpretation of the three regimes is currently based mainly on standardized anomalies of ERA5-derived predictors. I think the paper could be significantly strengthened by connecting these anomalies more directly with the actual meteorological environments.

In particular, I would encourage the authors to show, either in the main paper or in the Supplement, composites or representative values in physical units for the main thermodynamic and kinematic ingredients characterizing each regime. For example, CAPE, low-level and deep-layer shear, SRH, LCL and other key variables could be shown using their actual values rather than only standardized anomalies.

This would make the result much more physically tangible.

At present, Figure 6c gives the impression that regime 2 has a particularly strong signal in many of the physical predictors, whereas regimes 0 and 1 appear much less distinctive in this respect. This raises an interesting meteorological question: are the environments associated with regimes 0 and 1 genuinely much less characterized by these ingredients, or is part of this appearance simply a consequence of calculating anomalies relative to the same overall event population?

The paper would benefit from discussing this point more explicitly. It is particularly relevant because one of the potentially interesting contributions of the study is its possible connection with environmental prediction.

Response:

We thank the referee for this suggestion, which has substantially improved the physical interpretation. Numbers refer to the revised catalogue and to the two-regime solution now selected (General Response, points 1–3).

(i) Physical units: we now provide (a) **Table S4**, which gives the median and interquartile range of all 54 ERA5-derived parameters by regime, together with Cliff's δ and its 95 % interval obtained by resampling whole days, and δ for the last pre-event hour and for tornadoes only as robustness checks; (b) a new **Fig. S13**, which replaces the former anomaly profiles and shows boxplots in physical units of 12 parameters chosen on physical grounds (instability, inhibition, LCL, lapse rate, freezing level, low-level and deep-layer shear, SRH, 850 hPa wind, SHERB, WMAXSHEAR, STP) and the effect size of all 52 non-circular parameters; and (c) **Table S5**, with the share of events beyond reference values used in the literature. The methods are described in Sect. 2.5.5 ***(L349-363)*** and the results in Sect. 3.3 ***(L647-692)***. The table below summarizes the variables named by the referee together with the other ingredients of deep moist convection and tornadogenesis (Johns and Doswell, 1992; Doswell et al., 1996). These are the parameters shown in **Fig. S13a**, chosen on physical grounds rather than by the size of their differences (WMAXSHEAR 0–3 km, also shown there, is omitted here because it uses CAPE below 3 km only; see **Table S1**). All values refer to all events and to the mean over the 18 h before the event hour; **Table S4** gives the 54 parameters:

Parameter	Units	Regime 0: median [P25–P75]	Regime 1: median [P25–P75]	Cliff's δ [95 % CI]
ML CAPE	J kg^{-1}	122 [17–522]	56 [13–189]	-0.19 [-0.27, -0.10]
ML CIN	J kg^{-1}	-34 [-71–-17]	-23 [-41–-10]	0.23 [0.14, 0.32]
ML LCL height	m	942 [726–1208]	711 [608–921]	-0.36 [-0.45, -0.26]
Lapse rate 700-500 hPa	°C km^{-1}	6.3 [5.9–6.8]	6.5 [6.1–6.8]	0.09 [-0.01, 0.18]
0 °C isotherm height (MSL)	m	3248 [2402–3813]	2641 [2264–3105]	-0.29 [-0.36, -0.21]

Bulk shear 0-1 km	m s^{-1}	4.1 [2.9–5.8]	7.4 [4.2–10.1]	0.46 [0.37, 0.55]
Bulk shear 0-6 km	m s^{-1}	15.2 [11.2–21.3]	21.3 [16.5–26.3]	0.36 [0.28, 0.45]
SRH 0-1 km	m^2 s^{-2}	25 [8–50]	76 [29–129]	0.45 [0.36, 0.54]
Wind speed 850 hPa	m s^{-1}	6.3 [4.1–9.4]	12.2 [8.2–16.4]	0.53 [0.44, 0.61]
SHERB	–	0.55 [0.38–0.70]	0.78 [0.54–1.01]	0.42 [0.33, 0.51]
STP (fixed layer)	–	0.02 [0.00–0.11]	0.04 [0.01–0.15]	0.13 [0.03, 0.22]

The contrast is dominated by kinematics: of the 52 non-circular parameters, 13 have a medium or large effect ($|\delta| \geq 0.33$): eight wind-speed, shear and helicity diagnostics, two composites that include shear, and three low-level thermodynamic variables (lower LCL and LFC in regime 1). CAPE is lower in regime 1 (median MLCAPE 56 against 122 J kg^{-1}; $\delta = -0.19$), but it is low in both regimes (MLCAPE > 500 J kg^{-1} in 25 % of regime-0 and 5 % of regime-1 events), and the CAPE difference is negligible for the last pre-event hour ($\delta = -0.12$). Regime 1 is therefore a higher-shear, lower-CAPE, lower-cloud-base environment relative to regime 0, and 55 % of its events (against 24 % in regime 0) satisfy the usual definition of high-shear, low-CAPE environments (Sherburn and Parker, 2014): the regimes differ in frequency, not in kind. The same contrasts hold when only tornadoes are considered, with slightly larger effect sizes.

(ii) Anomalies relative to the same population: The referee's suspicion is correct in part, and it can be stated exactly. Anomalies are computed relative to the full sample, so for two regimes $z_0 = -(n_1/n_0)\, z_1$. With regime 0 containing 80 % of the events ($n_1/n_0 = 231/901 = 0.26$), an anomaly of +0.85 in regime 1 implies −0.22 in regime 0. The weak anomalies of regime 0 are thus a consequence of its size, not evidence that it is less characterized by these ingredients. Expressed as a standardized difference between the two regimes (the regime-1 anomaly divided by the share of events in regime 0), the leading predictors differ by about 0.85–1.07 standard deviations. The physical-unit values show that regime 0 is characterized by weaker shear, higher LCL and a higher freezing level, and by about twice the median MLCAPE of regime 1, although CAPE is low in absolute terms. This is now stated in Sect. 3.3 and in the caption of **Fig. 7c**, whose colour scale is fixed at ±0.2, ±0.5 and ±0.8 standard deviations.

(iii) Connection with environmental prediction. STP, SCP and EHI are close to zero in both regimes (STP ≥ 1 in about 1 % of events), whereas SHERB is higher in regime 1 (median 0.78 against 0.55; $\delta = 0.42$; ≥ 1 in 28 % against 6 % of events). Because the catalogue contains no non-tornadic cases, we cannot show that SHERB or any other parameter discriminates tornadic environments better; we have therefore removed this statement from the Discussion and Conclusions and present it as a hypothesis to be tested with non-tornadic events (Sect. 4, ***L833-839***; Sect. 5, ***L928-933***).

(iv) Caveats. Because the regimes were derived from these predictors, effect sizes describe the contrast rather than test it, and no significance tests are reported (Kriegeskorte et al., 2009; Sect. 2.5.5, ***L349-358***). Near-surface thermodynamic quantities depend on the 1000 hPa lowest level of the profiles (Sect. 2.3) and on ERA5 resolution (Taszarek et al., 2021).

7. The transition from occurrence frequency to high-threshold hazard requires stronger physical interpretation

One of the most interesting results is that regimes 0 and 1 dominate the overall tornado occurrence at moderate thresholds, while regime 2 becomes disproportionately important for the highest wind-speed exceedance.

This is an important and potentially powerful result. For that reason, I think it deserves a clearer meteorological interpretation. The reader would benefit from a discussion of why the environmental characteristics of regime 2 may favour the occurrence of the most intense reported tornadoes, and why the more frequent environments in regimes 0 and 1 do not necessarily produce the same high-end hazard.

The interpretation should be made carefully, given the limited number of high-intensity observations, but I think this is one of the places where the manuscript could make a stronger scientific contribution.

Response:

We thank the referee for this comment. With the revised catalogue this result has changed, and our interpretation with it. Two of the five EF3 reports of the discussion version were duplicates, and the former regime 2 is now part of regime 0. The share of EF2-or-stronger tornadoes is the same in both regimes (9.5 % in regime 0 and 9.6 % in regime 1; Fisher's exact test p = 1.0), and the three remaining EF3 tornadoes (all in regime 0, recorded in 1992–2003) do not constitute a statistically significant difference (p = 0.55). The lead of regime 0 at 180 and 220 km h^{-1} rests on these three events, whose large footprints dominate the exceedance at those thresholds, and at 180 km h^{-1} it reverses for windows starting in 2001 or later. We therefore no longer claim that a particular environment favours the most intense tornadoes, and the high-threshold contrast is explicitly presented as not robust (Sects. 3.4, 4 and 5).

The robust contrast is found at moderate intensities: the western–southwestern high-shear, low-CAPE regime has about 1.7 times the areal density of EF1-or-stronger tornadoes and the highest exceedance probability at 144 km h^{-1}, and its intensity distribution tends to be shifted from EF0 towards EF1, although this shift is partly due to outbreak clustering. The physical characterization requested in major comment 6 now allows a meteorological interpretation of both results (Sect. 4, ***L840-866***). First, this regime has the strongest low-level flow of the domain (median 850 hPa wind speed 12.2 against 6.3 m s^{-1}; Cliff's $\delta = 0.53$, the largest effect among the 52 non-circular parameters; **Table S4**). Strong ambient flow may add to the tornado's rotational winds and increase the damage produced by otherwise weak tornadoes (Wurman and Alexander, 2005), although differences in rating practice between regions cannot be excluded (Rodríguez et al., 2026). Second, significant tornadoes are typically associated with the joint presence of instability and vertical wind shear (Thompson et al., 2003). Neither regime combines the two: regime 1 has stronger shear but lower CAPE, and regime 0 the reverse. Their product differs little between the regimes (WMAXSHEAR 0–6 km: $\delta = -0.09$ for all events and -0.15 for tornadoes), and STP reaches 1 in less than 3 % of events in either regime (**Table S5**). This is consistent with the similar share of EF2-or-stronger tornadoes, although the small number of such events prevents a firmer conclusion.

8. The high-wind-speed results need careful treatment because of the very small EF3 sample

The authors correctly acknowledge that the 220 km h^{-1} results are based on only five EF3 reports and on the Brooks (2004) intensity–path scaling. However, because these results receive substantial attention in the paper, I think the limitations should be integrated even more strongly into their interpretation.

In particular, the statement around line 461 concerning the "cautious interpretation" is difficult to understand in its present form and should be rewritten more clearly.

The authors should also explain the shape of the curves in Figure 7a. Some of the apparent "bulges" within the different EF intervals look potentially artificial or at least non-intuitive. I would appreciate an explanation of whether these features arise from the mathematical formulation of the footprint model, the fitted distributions, the scaling between intensity classes, or another methodological aspect.

This would increase confidence in the exceedance curves.

Response:

We thank the referee for these comments.

(i) *EF3 sample:* the revised catalogue contains three EF3 tornadoes. Section 3.4 ***(L704-706, L712-716, L743-744)*** now states explicitly that the 220 km h^{-1} estimates rest on these three events, all recorded in 1992–2003; that they vanish for reporting windows starting in 2006 or later; and that the difference in EF3 share between regimes is not statistically significant. The dependence of the EF3 footprint on the Brooks (2004) path scaling is stated in Sect. 2.6 and among the limitations in Sect. 4 ***(L882-886)***. The sentence formerly at line 461 has been rewritten ***(L715-716)***: "This does not mean that EF3-like winds cannot occur in regime 1. It reflects the finite catalogue and the decision not to impose an EF3 rate where none was observed".

(ii) *Shape of the curves (**Fig. 7a** of the discussion version, now **Fig. 8a**):* the features noted by the referee arise from the formulation of the footprint fraction $f_k(\upsilon)$. Following the area-variability matrix of Ramsdell and Rishel (2007) and Balbi and Barbieri (2018), the footprint of each class is divided into fractions affected by each lower EF band, and within each band $f_k(\upsilon)$ decreases linearly from the lower to the upper wind-speed limit. The curves are

therefore continuous but change slope at the band limits (about 138, 179 and 219 km h^{-1}), where the contribution of a band ends, and the logarithmic scale emphasizes these changes. They are not produced by the Weibull fits or by the scaling between classes, which only set the expected footprint area E [A_k], a constant factor for each class. This is now explained in Sects. 2.6 and 3.4 ***(L404-406, L716-719)***.

Other comments and specific points

9. Definitions of QC1 and QC2

QC1 and QC2 should be defined at their first occurrence in Section 2.1. Readers who are not familiar with the ESWD quality-control terminology should not have to search for the definitions elsewhere.

Response:

We have added the definitions of the ESWD QC1 and QC2 quality-control levels to Sect. 2.1 ***(L105-107)***: QC1 denotes a report confirmed by a reliable source, and QC2 an event fully verified, for example through a scientific case study (Dotzek et al., 2009). Reports at both levels were retained, whereas unverified reports (QC0 and QC0+) were excluded.

10. Clarify the term "filtering"

The manuscript repeatedly refers to "filtering" of the database, but this is not sufficiently specific. Please state exactly which filters are applied and in what order.

Response:

Section 2.1 ***(L104-112)*** now lists the filters explicitly and in the order applied: (i) retention of reliable reports only (SINOBAS reports of high or medium reliability and ESWD reports at QC1 and QC2); (ii) retention of tornadoes and waterspouts only, excluding other vortex-like or convective-wind phenomena; (iii) restriction to peninsular Spain and the Balearic Islands; (iv) restriction to 1992–2025; (v) harmonization of event times to UTC; and (vi) removal of duplicate reports, which is performed last and described in detail in the same section and in the new Fig. S1.

11. Time convention: UTC versus local time

Section 2.1 states that event times were harmonized to UTC, while Section 2.2 discusses climatological distributions in local time. Although the difference may not be very large in this geographical domain, the distinction should be made completely clear and the terminology should be consistent. I would personally recommend using UTC as the standard time convention throughout the methodological description, with an explicit conversion to LT only where the diurnal climatology is discussed.

Response:

Times were harmonized to UTC for duplicate detection and ERA5 extraction, as stated in Sect. 2.1, and UTC is used throughout the methods. The diurnal distribution (**Fig. 3c**, former **Fig. 2c**) is presented in local time (LT; standard or daylight-saving time) for physical interpretability; this convention is now defined once in Sect. 2.2 ***(L151-152)*** and used consistently in the text introducing the diurnal cycle (Sect. 3, ***L499-502***) and in the caption of **Fig. 3**.

12. Figure 1

It would be useful to include latitude and longitude axes in Figure 1. This would make the geographical limits of the proposed regions easier to interpret.

The sentence in the caption describing how the map was generated with GeoPandas–Matplotlib and the geographical base layers may also be reconsidered. If retained, it should be clear why this technical information is useful to the reader.

Response:

We have added latitude and longitude labels, with units, to **Fig. 2 (former Fig. 1)** and updated the event counts in its legend to the revised catalogue. The technical sentence on how the map was generated has been removed from the caption; the software used is described in Sect. 2.7 and in the Code availability section.

13. Clarify the "hourly series" and selected times

Around line 137, the manuscript refers to hourly series but then summarizes them at t = −6, 0 and +6 h. Please clarify whether the complete hourly series are used in the analysis and these three times are only selected for illustration, or whether the analysis is actually based on these selected times and the other descriptors.

Response:

The complete hourly series are used. For each event, hourly profiles were extracted from 18 h before to 6 h after the event hour, and all analyses use only the 18 pre-event hours (t = −18 to −1 h). Each series is summarized by eight descriptors: the value at t = −6 h, the mean, maximum, minimum, the last pre-event value, the change between the first and last pre-event hours, the standard deviation and the linear slope. The value at t = −6 h is therefore one of these descriptors, not a selection of times, and values from the event hour onward (t = 0 to +6 h) are not used in any analysis. This is now stated explicitly in Sect. 2.3 (***L186-192***).

14. Clarify the PCA dimensionality in Section 3.2

There appears to be some inconsistency, or at least potential confusion, between the statement that up to 10 PCs are retained to explain approximately 90 % of the variance and the fact that the discussion in Section 3.2 focuses on the first two PCs.

Please explain explicitly how many PCs are used in the actual clustering/classification analyses and why Figures 3a and the associated discussion focus only on PC1 and PC2.

Response:

We thank the referee for identifying this inconsistency. Numbers below refer to the revised 1132-event catalogue and the corrected analysis (General Response, points 1–2). The statement in the original Sect. 2.4 ("≈ 90 % of the total variance (up to 10 components)") did not describe the procedure actually applied and has been corrected. Section 2.4 has been reorganized (see our response to major comment 4), and the dimensionality used in each analysis is now stated explicitly in the new Sect. 2.4.1 (***L211-218***). Variance is spread over many components in this 432-predictor space: the first 2, 10 and 30 PCs explain 30.0 %, 58.7 % and 79.8 % of the total variance, respectively **(new Fig. S3**, reproduced below).

(i) In the supervised classification, the number of PCs was treated as a tunable setting (5, 10, 15, 20 or 30) and selected within the inner loop of the nested, spatially grouped cross-validation, with the PCA fitted on the training folds only. The best classifier (SVM-RBF) retained 30 PCs (79.8 % of the variance) in four of the five outer folds, and the numbers most frequently selected by the other classifiers were 20 PCs for logistic regression (72.4 %), 15 for the decision tree (66.7 %) and 10 for Gaussian naive Bayes (58.7 %).

(ii) In the unsupervised screening, DBSCAN and Ward candidates were computed with 5–30 leading PCs, whereas the grid-based diagnostic operates on the PC1–PC2 plane by construction.

(iii) In the spatial regionalization used for the hazard analysis, PCA was applied to the cell-level matrix (median of the events' predictors in each cell); with only about one hundred cells, the number of components was capped at 15, which explain 77.9 % of the variance at 50 km. The retained Ward partition (50 km, k = 2) is computed from these 15 leading PCs (Sects. 2.5.1–2.5.2, ***L295-299***).

(iv) **Figure 4a** shows only PC1 and PC2 because a two-dimensional projection is the only practical way to display the event cloud. Since these two components explain 30 % of the variance, the panel is now explicitly presented as a

qualitative illustration, and the separability conclusions rest on the classification performed in the higher-dimensional PC space **(Fig. 4b, c)**.

The cap of 30 components lies within the 70–90 % cumulative-variance range commonly used as a guideline (Jolliffe, 2002) and limits dimensionality for distance-based methods (Beyer et al., 1999), while selecting the number of PCs inside the cross-validation avoids optimistic bias (Hastie et al., 2009; Cawley and Talbot, 2010).

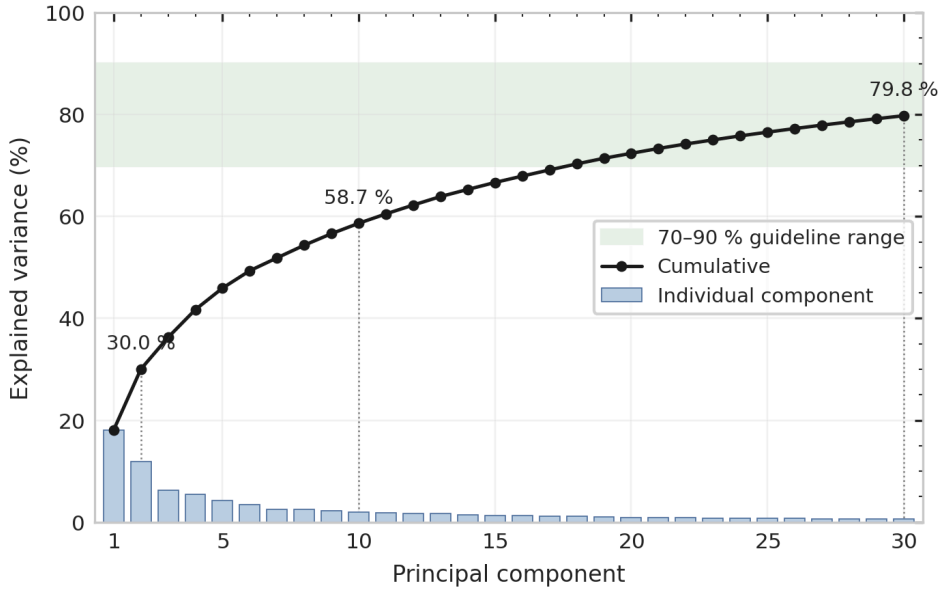


Figure S3. Variance explained by the principal components of the event-level predictor matrix (1132 events, 432 standardized predictors). Bars show the variance explained by each component and the line the cumulative variance; the first 2, 10 and 30 components explain 30.0 %, 58.7 % and 79.8 % of the total variance. The shaded band marks the 70–90 % range commonly used as a guideline (Jolliffe, 2002); the number of retained components was capped at 30 (Sect. 2.4.1).

15. Figure 3: explained variance

Please indicate in Figure 3, or in its caption/text, what percentage of the total variance is explained by PC1 and PC2. This is important for interpreting how representative the two-dimensional projection is.

Response:

We thank the referee for this suggestion. Numbers below refer to the revised 1132-event catalogue and the corrected analysis (General Response, points 1–2). The percentages of total variance explained by the first two principal components are now reported in the axis labels of **Fig. 3a** of the discussion version **(now Fig. 4a)** (PC1: 18.1 %; PC2: 12.0 %), in the figure caption and in Sect. 3.2 ***(L529-535)***. Together they explain 30 % of the total variance. This is expected in a high-dimensional predictor space combining thermodynamic, kinematic and composite diagnostics, in which variance is distributed over many modes (Jolliffe and Cadima, 2016; Wilks, 2019). Both components nevertheless carry a clear signal: they exceed the broken-stick expectation by about an order of magnitude (Jackson, 1993) and are well separated according to the sampling-error criterion of North et al. (1982). The text now states that **Fig. 4a** is a qualitative illustration and that the separability conclusions rest on the supervised classification performed with up to 30 PCs (79.8 % of the variance; new Sect. 2.4.1 and our response to comment 14).

16. Rejection of DBSCAN and other candidate methods

The final part of Section 3.2, where DBSCAN and the other approaches are rejected in favour of Ward clustering, would be easier to understand if the authors showed some details or a simple example of the candidate solutions.

For example, one additional panel or a concise table showing the main candidate solutions and why they were rejected would make the methodological decision much more transparent.

Response:

We agree that this decision needed to be more transparent. Numbers below refer to the revised 1132-event catalogue and the corrected analysis (General Response, points 1–2). We made two changes. First, Sect. 2.4.4 ***(L265-275)*** now defines the noise fraction and the dominant-cluster fraction with a simple numerical example and states the screening thresholds. Second, the screening table (**Table S3**) now reports, for each method, the number of tested and valid configurations, the best raw candidate, its noise and dominant-cluster fractions and the reason for rejection, with the thresholds in the caption, and **Fig. S6** shows these diagnostics graphically. Neither DBSCAN (0 of 90 configurations) nor the grid-based method (0 of 20) produced a valid candidate: their best raw candidates placed at least 99 % of the clustered events in a single cluster. In contrast, 29 of the 30 Ward configurations passed the screening.

17. Clarify the last paragraph of Section 2.6

The final paragraph of Section 2.6, particularly its last two sentences, is difficult to read and assimilate. I recommend rewriting this paragraph using shorter sentences and making the sequence of the occurrence, footprint and exceedance calculations more explicit.

Response:

We agree. The paragraph has been rewritten with shorter sentences (Sect. 2.6, ***L365-369***). It now defines, in sequence, the point-scale exceedance probability, the neighbourhood return period and how its rate is computed, and then explains why the two metrics can rank regimes differently.

18. Length and width notation

At line 231, the notation "L × W" is introduced. Please define explicitly that L and W correspond to path length and mean/path width, respectively, and check that the terminology is used consistently throughout the manuscript.

Response: We have made this explicit: L and W are now defined at their first use as path length and mean path width (Sect. 2.6, ***L384***), and the same terms are used consistently throughout the manuscript.

19. Figure 8 and the presentation of hazard

For hazard communication, return periods are often more tangible and easier to interpret than very small exceedance probabilities. Since the manuscript already calculates return periods, I suggest considering whether Figure 8 could also be presented, or complemented, in terms of return periods.

This could make the spatial results more useful for readers interested in hazard assessment and applications.

Response:

We agree that return periods are easier to communicate. A new supplementary figure (**Fig. S19**) maps the 50 km neighbourhood return period of each cell at 144 and 180 km h^{-1}, computed from the same smoothed cell rates as **Fig. 8** of the discussion version (**now Fig. 9**); the 220 km h^{-1} threshold is not mapped because it rests on three EF3 tornadoes. The median cell return period is 6.9 years in regime 1 against 14.1 years in regime 0 at 144 km h^{-1}, and 77 against 197 years at 180 km h^{-1} (Sect. 3.4, ***L772-775***). Regime-level return periods with bootstrap intervals remain in **Fig. 8b** and **Table S6**. To keep the regime-level figure legible at page width, the former fixed-threshold panel, which repeated the values shown by the curves in panel (a) and quoted in Sect. 3.4, has been removed. We kept the exceedance probability in **Fig. 9** because it is the quantity produced by the footprint model and the one used in design-basis assessments (Ramsdell and Rishel, 2007), and present the return period as a complement.

20. Potential exposure/socioeconomic perspective

The final part of Section 3.4 could potentially be strengthened by comparing the hazard climatology with population density or another simple socioeconomic/exposure indicator, using the same hexagonal grid adopted in the study.

I consider this an optional suggestion rather than a requirement, but such a comparison could considerably increase the practical relevance of the work by showing where the identified hazard overlaps with areas of greater exposure.

Response:

We thank the referee for this suggestion. As part of our response to RC2 (major comment 2), we assigned the Census 2021 1 km² population grid (Eurostat, 2026) to the same 50 km hexagonal cells used in the hazard analysis (45.2 million residents in the study domain) and compared it with the tornado rates (**new Fig. S16**; Sect. 3.4, ***L749-759***). Three results bear on exposure. First, the regime with the highest moderate-threshold hazard (regime 1) is about as densely populated as regime 0 (96 against 88 persons km^{-2}) and is home to about 14 million people, roughly a third of the population of the domain. Second, the cells with the highest exceedance probabilities at 144 km h^{-1}, along the coasts of the Gulf of Cádiz, Valencia, Catalonia, Galicia and Mallorca (**Fig. 9**), are among the most densely populated (**Fig. S16a**), so the hazard and the population largely overlap. Third, this overlap is partly circular: report rates increase by a factor of 1.64 per doubling of population density in both regimes, so sparsely populated cells are probably underestimated, although part of this association is climatological, since coastal cells are both densely populated and favourable for tornadoes. Controlling for density changes the regime contrast by only 5–6 %, so the contrast does not arise from differences in population. A full exposure analysis, combining population, buildings and critical infrastructure at the cell level, requires data and methods beyond the scope of this study, and we now identify it in the Conclusions as a natural next step.

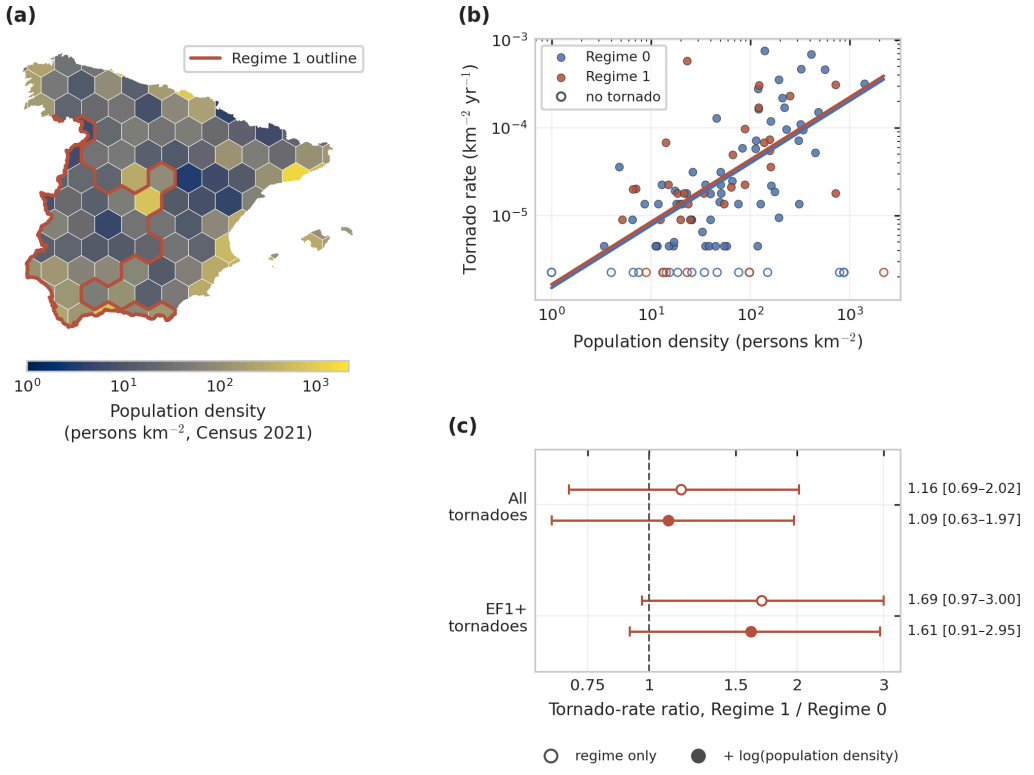


Figure S16. Sensitivity of the regime contrast in tornado occurrence to population density, used as a proxy for reporting effort. (a) Population density of each hexagon from the Census 2021 1 km² population grid (Eurostat GEOSTAT; INE for Spain), logarithmic colour scale, with the outline of regime 1. (b) Hexagon tornado rate against population density (log-log axes); hollow markers at the bottom are hexagons without tornadoes, and the lines are the fitted Poisson model with a common density effect, so their vertical offset is the regime contrast after controlling for population. (c) Tornado-rate

ratio of regime 1 relative to regime 0 for all tornadoes and for EF1-or-stronger tornadoes, without (open) and with (filled) the population-density control; whiskers are 95 % intervals from 2000 bootstrap resamples of hexagons within regime.

21. Conclusions

The Conclusions are concise and acceptable, but they largely repeat points already presented in the Discussion. The authors could consider making this section slightly more focused on the main new findings and implications rather than repeating the preceding discussion.

Response:

We agree. The Conclusions have been rewritten around four main findings, ordered as the objectives of the study, and three implications, avoiding repetition of the Discussion (Sect. 5, ***L905-946***).

22. Supplementary Figure S7

I recommend carefully checking and comparing the anomalies shown in Figure S7 with those presented in Figure 6c. At first sight, there appear to be regime-dependent differences, including some apparently discrepant signs.

This should be checked carefully because it may affect the desired meteorological interpretation of the three regimes. If the two figures use different variables, rankings, normalizations or samples, this should be stated very clearly. If the two figures are intended to represent the same quantities, the apparent differences should be resolved.

Response:

We thank the referee for this careful check. Both figures were computed from the same table of standardized anomalies (same predictors, same event sample, anomalies relative to the full-sample mean), but with different selection rules: **Fig. 6c** of the discussion version (**now Fig. 7c**) showed the 15 predictors with the largest absolute anomaly in any regime, a ranking dominated by the regime with the largest departures, whereas **Fig. S7 (now Fig. S13)** showed the 20 leading predictors of each regime separately. Different predictors therefore appeared in each figure, which gave the impression of discrepant signs; moreover, with three regimes, a predictor can legitimately have opposite signs in different regimes. With the two-regime partition now retained (General Response, point 3), this ambiguity disappears: because anomalies are computed relative to the full-sample mean, the anomalies of the two regimes satisfy $z_0 = -(n_1/n_0)\, z_1$, so they have opposite signs, magnitudes inversely proportional to the regime sizes, and the same ranking. We verified that this relation holds for all 432 predictors, which confirms that the anomalies of both regimes were computed from the same sample, mean and standard deviation. **Fig. 7c** is now the only figure showing standardized anomalies, and its caption states the selection rule (***L641-646***). The former anomaly profiles of **Fig. S13** have been replaced by the regime contrast in physical units and effect sizes, and values in physical units are given in **Table S4** (see our response to major comment 6).

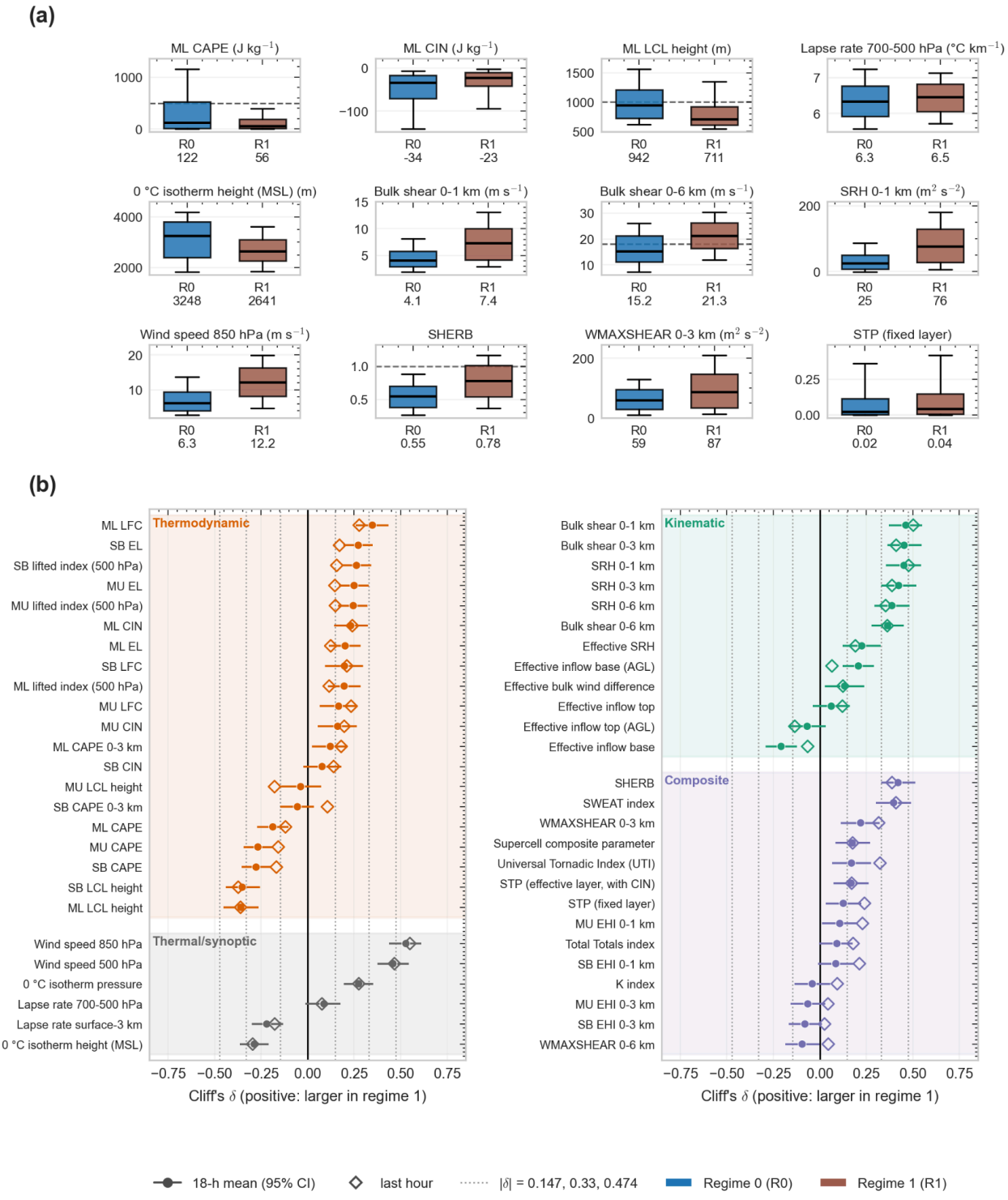


Figure S13. Regime contrast in physical units. (a) Distributions of 12 parameters chosen on physical grounds (all events; mean over the 18 h before the event hour): boxes span the 25th–75th percentiles, the line is the median and whiskers the 10th and 90th percentiles; the numbers under each regime label are the medians. Dashed horizontal lines mark reference values (MLCAPE 500 J kg^{-1}, MLLCL 1000 m, 0–6 km bulk shear 18 m s^{-1}, SHERB 1) and are drawn only when they fall within the plotted range; the reference value of STP (1) lies above the 90th percentile and is not shown (Table S5). WMAXSHEAR 0–3 km uses surface-based CAPE below 3 km (Table S1). (b) Cliff's δ between regime 1 and regime 0 for

the 52 non-circular parameters, grouped by category. Dots: 18-h mean with 95 % bootstrap intervals (resampling whole days); diamonds: last pre-event hour. Positive values indicate larger values in regime 1. Dotted vertical lines mark |δ| = 0.147, 0.33 and 0.474 (Romano et al., 2006). Regimes were derived from the same predictors, so the figure describes the contrast and does not test it.

Citation: *https://doi.org/10.5194/egusphere-2026-4053-RC1*

Reviewer 2:

General assessment

The manuscript is a very interesting contribution to the study of tornadoes and waterspouts in Spain and also in Europe. The results are relevant, but in the current form the manuscript is not ready for publication.

Response:

We thank the referee for the positive assessment and the constructive comments. All comments have been addressed; the main changes affecting several responses (revised catalogue, code corrections, selected regionalization and reorganized Methods) are summarized in the General Response above.

Major comments

1. The Methodology section is very dense and with no sufficient details. In the current version of the manuscript it reads as a collection of statistical analyses without a clear structure. I recommend to the authors to reorganise the Methodology section so that all the relevant details, the purpose of each part of the analysis and the link to the manuscript objectives are explicit.

Response:

We thank the referee for this comment, which coincides with the main concern of RC1. The Methods have been reorganized so that the purpose of each step and its link to the objectives are explicit. A new introductory paragraph and workflow figure **(Fig. 1)** link each step to objectives 1–4 and to its results section; Sects. 2.4 and 2.5 now open with the question they address and are divided into short subsections in the same order as the results; and the complete settings are given in the Supplement (**Table S2, Sect. S1**). Details are given in our response to RC1, major comment 4.

2. The authors mention that the rising number of reports over the study period is probably due to better reporting and increase public awareness and not due to more tornadoes. I agree with this observation, because this is the situation in other European countries as well regarding severe weather reports in general (not just tornadoes). Then the authors have used the data to calculate tornado rates assuming that the reporting is homogeneous both in space and time. This is an important point, given that regime 2 covers the sparsely populated central-southeastern Spain, where reporting is likely low (and only big damaging events are reported). The authors can conduct a sensitivity analysis using recent years (e.g., after 2005), and normalise by a reporting proxy (e.g., population, road density). This can show whether the regime ranking is robust or partly a reporting artefact.

Response:

We thank the referee for this important point. In the revised regionalization, the former regime 2 is part of regime 0 (General Response, point 3). We carried out two complementary sensitivity analyses (Sects. 2.6 and 3.4; Supplement, **Fig. S15**, reproduced below, and **Fig. S16**, shown in our response to RC1, comment 20).

(1) *Reporting period:* rather than choosing arbitrary start years, we anchored the windows on documented changes in the two reporting systems that feed the catalogue and on a data-driven check. The occurrence rates were recomputed, with the footprint model held fixed, for windows ending in 2025 and starting in 2001 (the change point in the annual number of reports detected by the Pettitt (1979) test, $p = 0.006$, with the mean rising from 8.8 to 18.4 reports per year), 2006 (start of operations of the ESWD, the referee's "after 2005"; Groenemeijer and Kühne, 2014; Groenemeijer et al., 2017) and 2013 (launch of SINOBAS). At 144 km h^{-1}, regime 1 has the shorter return period in the full record and in each of the three shorter windows, and the bootstrap intervals of the two regimes never overlap. At 180 km h^{-1}, regime 0 leads in point-scale exceedance only in the full record, because of three EF3 tornadoes recorded in 1992, 1999 and 2003; regime 1 leads in every window starting in 2001 or later, but the intervals overlap because few EF2–EF3 tornadoes remain. At 220 km h^{-1}, no estimate is available from 2006 onwards. The analysis also revealed a change in rating practice: the share of unrated reports rose from 25 % to 51 % in regime 0 and from 17 % to 28 % in regime 1 between the full record and 2013–2025. This partly explains why the regime contrast widens in recent windows, so we report the full-record contrast as the conservative estimate. The

high-threshold contrast is therefore conditional on the early part of the record, and it is now presented as such in the text.

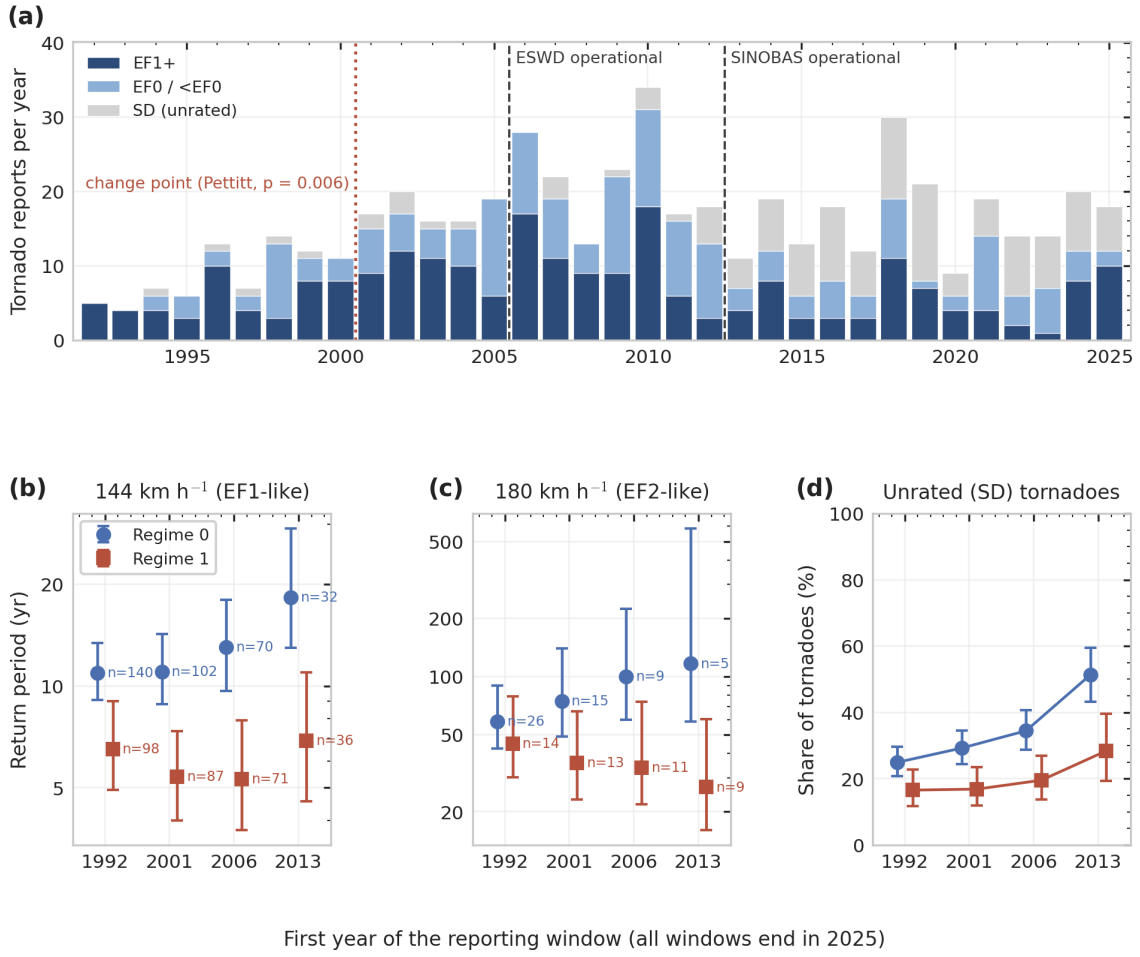


Figure S15. Sensitivity of the regime contrast to the reporting period. (a) Annual number of tornado reports by intensity group, 1992–2025; dashed lines mark the start of operations of the two reporting systems that feed the catalogue (ESWD, 2006; SINOBAS, April 2013). The Pettitt change point of the annual number of reports (dotted line) is 2001 (p = 0.006). (b, c) Bootstrap median return period of threshold-capable tornadoes within 50 km at (b) 144 and (c) 180 km h^{-1}, for windows starting in 1992, 2001, 2006 and 2013 and ending in 2025; whiskers are 5th-95th percentiles from 5000 resamples of years, and n is the number of threshold-capable tornadoes. The footprint model is held fixed, so only the occurrence rates change between windows. (d) Share of tornadoes without an intensity rating (SD) in each window and regime, with 95 % Wilson intervals.

(2) *Reporting density:* cell-level counts of all tornadoes and of EF1-or-stronger tornadoes (the classes that control exceedance at 144 km h^{-1}) were modelled with a Poisson regression on regime and log population density from the 1 km² Census 2021 population grid (Eurostat, 2026), with the cell area A_h times the number of years T as exposure. Because the counts are overdispersed (Pearson dispersion 8.5 and 5.3), inference was based on quasi-Poisson tests and on a bootstrap of cells within regimes (McCullagh and Nelder, 1989). Population density is the dominant predictor of report rates (rate ratio 1.64 and 1.51 per doubling; $p < 10^{-6}$), confirming, as the referee anticipated, that reporting effort shapes the absolute rates. However, the two regimes are similarly populated (96 and 88 persons km^{-2}), and the population control changes the regime contrast by only 5–6 %. For all tornadoes the rate ratio of regime 1 to regime 0 is close to 1 (1.16 without and 1.09 with the control; p = 0.74). For EF1-or-stronger tornadoes

it is 1.69 without and 1.61 with the control (95 % interval 0.91–2.95; 95 % of bootstrap replicates above 1; p = 0.13): the contrast is not statistically significant at the cell level, where counts are few and clustered, but it is not explained by population. These results are reported in Sect. 3.4 ***(L749-759)*** and discussed as a limitation in Sect. 4 ***(L880-882)***.

In both tests, the ranking of the regimes at moderate thresholds is unchanged, so it is not explained by changes in reporting over time or by differences in population density, whereas the high-threshold contrast depends on the early part of the record.

3. The labels for the regimes are obtained from medians at the hexagon level of the pre-event predictors derived from ERA5. The classifier is then used to recover the labels from event values of the same predictors. But the hexagon is already a summary of the events. This seems like a circular argument, or is just the methodology that is not described clearly. One suggestion here is to repeat the Ward clustering keeping out each hexagon, relabel the hexagon and use the classifier on those independently obtained labels.

Response:

We thank the referee for raising this point, which we consider important. Numbers below refer to the revised catalogue and to the two-regime solution now selected (General Response, points 1–3). The referee is correct that the separability test is not fully independent: the regime labels derive from cell medians of the same predictors used by the classifier, and each median includes the events later predicted. Spatial blocking limits but does not remove this dependence. We now state this explicitly (Sect. 2.5.3, ***L320-326***) and interpret the scores as measures of internal consistency and spatial transferability rather than as external validation.

We also implemented the referee's suggestion. Each of the 102 occupied cells was left out in turn, the constrained Ward partition was recomputed without it, an SVM-RBF classifier was trained on the events of the remaining cells with their new labels, and the left-out cell received the regime predicted for the majority of its events. In this test, 85 of the 102 cells (83 %) recovered their original regime (Cohen's κ = 0.57; ARI = 0.41), 89 % for regime 0 and 67 % for regime 1. Disagreements were more frequent in cells adjacent to the regime boundary (29 %) than in interior cells (12 %; Fisher's exact test, p = 0.07), and cells that changed regime contained a similar number of events to those that did not (median 4 against 5). We conclude that the regimes are largely reproducible from independently derived labels, but that the assignment of individual cells, particularly near the boundary, carries some uncertainty; the text now describes the boundary accordingly (Sect. 3.3, ***L628-635***; **Fig. S12**). For computational reasons, the classifier in this test used fixed settings (C = 1, γ = "scale") instead of the nested grid search.

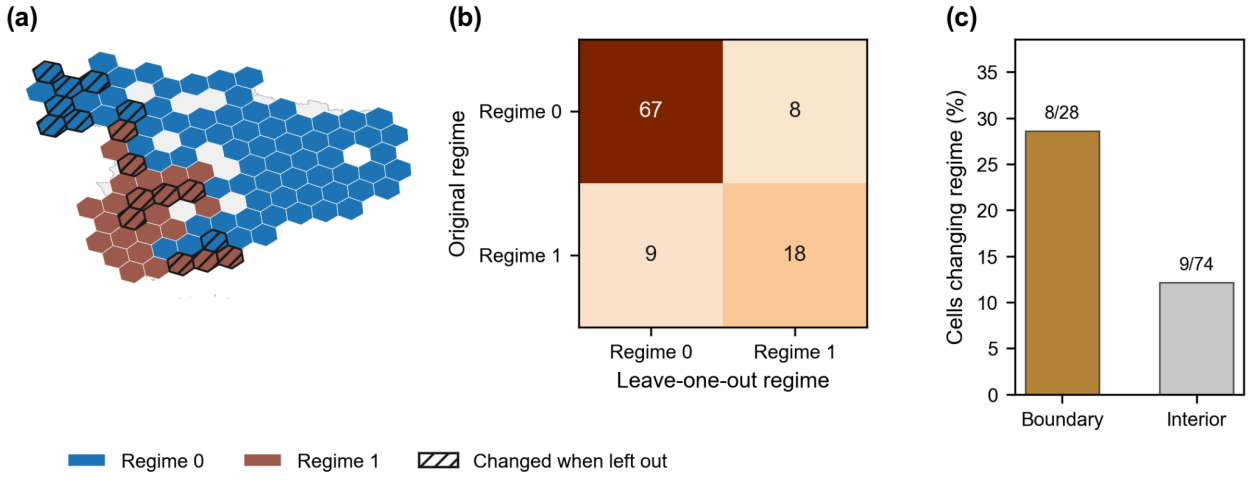


Figure S12. Leave-one-cell-out test of the regionalization. Each of the 102 occupied 50 km cells was removed in turn, the constrained Ward partition (k = 2) was recomputed without it, an SVM-RBF classifier with fixed settings (C = 1, γ = scale) was trained on the events of the remaining cells with their new labels, and the removed cell received the regime

predicted for the majority of its events. (a) Cells coloured by their original regime; hatched cells were assigned the other regime when left out. (b) Original versus leave-one-out regime (83 % agreement; Cohen's κ = 0.57; adjusted Rand index 0.41). (c) Share of cells whose regime changed, for cells adjacent to the regime boundary (8 of 28) and interior cells (9 of 74); Fisher's exact test p = 0.07.

4. One of the main points of the manuscript is that the least active regions are also the regions where the stronger winds are more likely. This is based on 4 (out of 5) EF3 tornadoes from regime 2. This number of events is very small. The authors should provide a table with the number of tornadoes by EF class and regime. Also, the authors should test whether the intensity distribution is different between different regimes.

Response:

We thank the referee for this suggestion. A new table in the main text (**Table 2**) gives the number of tornadoes by EF class and regime. We tested whether the intensity distributions differ between regimes (Sect. 2.6; Supplement, **Fig. S14**). For the 420 rated tornadoes, the distributions differ (permutation χ^2 = 13.2, 200 000 replicates, p = 0.008; Cramér's V = 0.18), but the difference weakens when only one tornado per day is kept to remove outbreak clustering (median p = 0.04; p < 0.05 in 62 % of subsamples). The standardized residuals show that the difference lies between EF0 and EF1 (largest residuals in these classes, |r| = 1.3–1.8, against ≤ 0.3 for EF2). In contrast, the share of EF2-or-stronger tornadoes is almost identical: 9.6 % in regime 1 (95 % Wilson interval 5.8–15.5 %) against 9.5 % in regime 0 (6.6–13.5 %), and the EF3 difference (3 of 274 against 0 of 146) is not significant (p = 0.55): the sample is too small to distinguish the regimes at the high end. The referee's concern is therefore confirmed: the EF3-based contrast of the discussion version is not statistically supported, and the revised text no longer claims it (Abstract; Sects. 3.4, 4 and 5).

Table 2: Number of tornadoes assigned to each regime, by intensity class, 1992–2025.

Regime	SD	<EF0	EF0	EF1	EF2	EF3	Total	Area (km²)
0	91	7	127	114	23	3	365	352 881
1	29	1	47	84	14	0	175	145 749

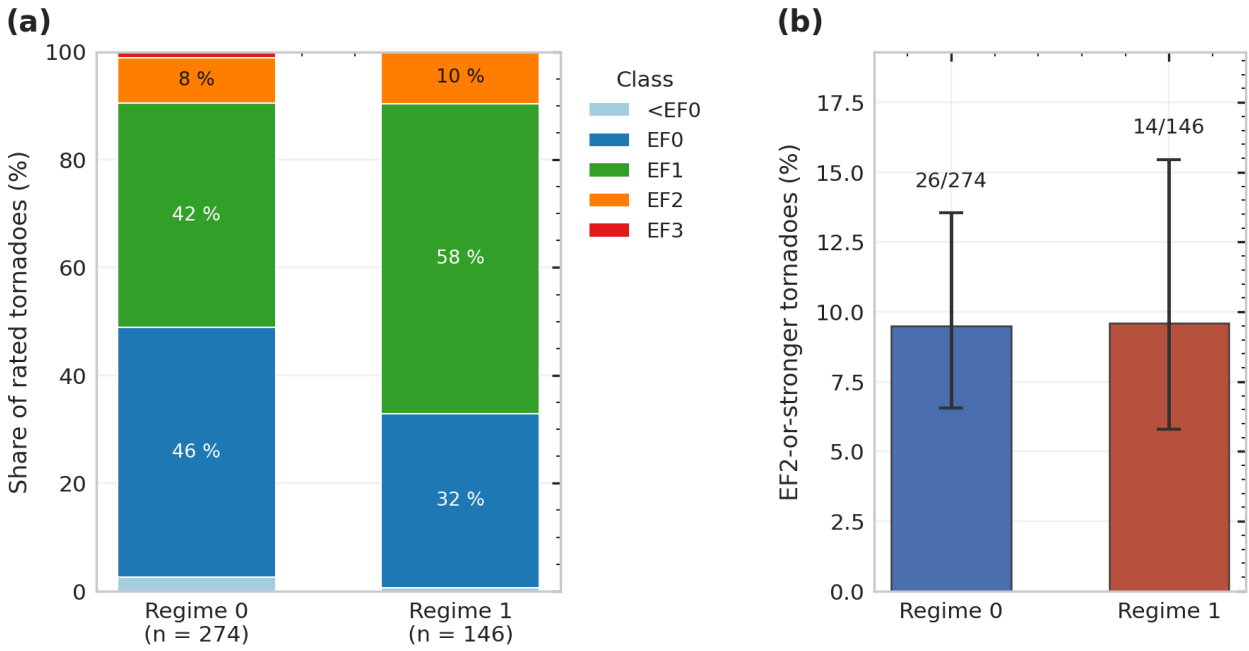


Figure S14. Intensity distribution of the rated tornadoes by regime. (a) Share of each EF class (n = 420 rated tornadoes). The distributions differ (permutation chi-squared test, 200 000 replicates: $\chi^2 = 13.2$, $p = 0.008$, Cramér's V = 0.18; Good, 2005), but the difference weakens when only one tornado per day is kept (median $p = 0.04$; $p < 0.05$ in 62 % of 200 subsamples). (b) Share of EF2-or-stronger tornadoes with 95 % Wilson intervals (Wilson, 1927); labels give the number of EF2+ tornadoes over the number of rated tornadoes. Fisher's exact test: EF2+ share p = 1.00; EF3 share (3 of 274 against 0 of 146) p = 0.55.

5. Waterspouts represent approximately 50% of the events in the dataset and are concentrated in regime 0. Thus, the regime 0 "weakly forced, low-shear" signature may be a marine boundary-layer signature rather than a tornadic environment. This is then used as the spatial frame for a tornado only hazard calculation. The Ward regionalization should be done also on tornadoes alone to show that the partition is stable. Also, some of the hexagons had no reports and were labelled by neighbour inference (Figure 5b). These hexagons contribute to the regime areas A_r used in equation 1 affecting the occurrence rate. The authors should calculate the rates based on observed hexagons only and include this in the manuscript as well.

Response:

We thank the referee for these suggestions. Numbers below refer to the revised catalogue and to the two-regime solution now selected (General Response, points 1–3). In the current partition, the former regime 2 (central–southeastern belt) is part of regime 0.

(1) *Tornado-only regionalization:* we repeated the full regionalization procedure (50 km cells, k = 2, same PCA and constrained Ward settings) using tornadoes only (540 events, 91 occupied cells). Comparing the regime assigned to each tornado by the two partitions, 96.7 % of the 492 tornadoes located inside the clipped domain fall in the same regime (adjusted Rand index = 0.87; normalized mutual information = 0.78; Strehl and Ghosh, 2002; **Fig. S11**). The partition is therefore essentially unchanged when waterspouts are excluded. Moreover, the physical-unit characterization (**Table S4**) shows that the regime contrast is also present when only tornadoes are considered: regime-1 tornadoes form with stronger 0–1 km bulk shear (Cliff's $\delta = 0.51$; medians 8.3 against 4.5 m s^{-1}) and lower CAPE (MLCAPE $\delta = -0.27$; medians 73 against 208 J kg^{-1}) than regime-0 tornadoes, with effect sizes slightly larger than for all events. The signature is therefore not merely a marine boundary-layer signature of waterspouts. This test is described in Sect. 2.5.3 and reported in Sect. 3.3 (***L624-628***). This is also stated in the Discussion ***(L825-827)***.

(2) *Rates from observed cells only:* we recomputed the occurrence rates using only cells with at least one observed event. The regime areas decrease by 13.5 % (regime 0) and 11.2 % (regime 1), and the exceedance probabilities increase accordingly, by 16 % and 13 %, while the ranking of the regimes is unchanged at every threshold (Sect. 3.4, ***L759-763***; **Fig. S17**). Because the rate is inversely proportional to the regime area (Eq. 1), this is a uniform rescaling of each regime, and its size is smaller than the sampling uncertainty of the estimates (e.g. the 5th–95th percentile range of the return period at 144 km h^{-1} spans 4.9–9.0 years in regime 1 and 9.1–13.4 years in regime 0; **Table S6**). The two estimates bracket the true value: the observed-cell rates are an upper bound if cells without reports truly had fewer tornadoes, and the full-area rates a lower bound if those cells are under-reported.

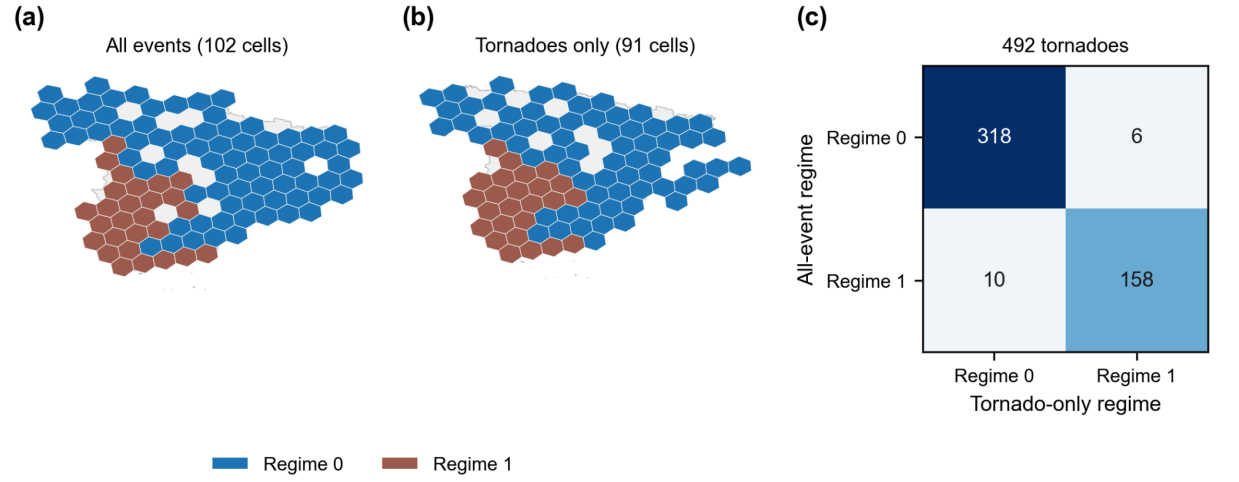


Figure S11. Robustness of the regionalization to the exclusion of waterspouts. (a) Two-regime partition of the occupied 50 km cells derived from the combined tornado and waterspout sample. (b) Partition obtained with the same procedure applied to tornadoes only (540 events, 91 occupied cells); its regime labels were matched to those of panel (a) by maximum overlap. (c) Number of tornadoes assigned to each regime by the two partitions, for the 492 tornadoes located inside the clipped domain: 96.7 % fall in the same regime (adjusted Rand index 0.87; normalized mutual information 0.78).

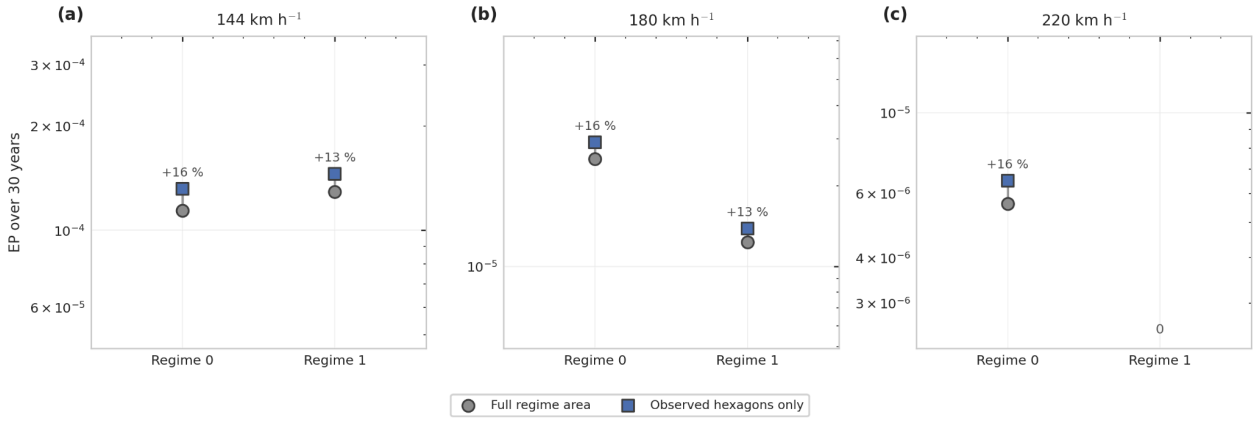


Figure S17. Fixed-threshold 30-year tornado exceedance probability using the full (neighbour-inferred) regime area (grey circles, main analysis) versus using only directly observed hexagons (blue squares) for A_r in Eq. 1; the segment joins the two estimates. The footprint/exceedance model is unchanged; only the exposure area differs. The regime ranking is identical at every threshold.

Minor comments

1. The ESWD data are not described in details, for example what are QC1, QC2?

Response:

We have added the definitions of the ESWD quality-control levels to Sect. 2.1 (***L105-107***): QC1 denotes a report confirmed by a reliable source and QC2 an event fully verified, e.g. through a scientific case study (Dotzek et al., 2009). Reports at both levels were retained, whereas unverified reports (QC0 and QC0+) were excluded.

2. For the analysis based on ERA5 reanalysis, how many vertical levels were used for the pseudo-soundings? Are this model levels?

Response:

We used the 37 ERA5 pressure levels (1000–1 hPa) rather than model levels, and this is now stated in Sect. 2.3 ***(L171-175)***. Geopotential, temperature, relative humidity and wind were extracted at the four grid points enclosing each event, the 54 parameters were computed from each profile with SHARPpy (Blumberg et al., 2017), and the resulting values were averaged over the four points. We now also state a limitation of this choice. Since no surface fields were used, the lowest level of each profile is 1000 hPa, which lies below the ground over elevated terrain and is extrapolated in ERA5. Surface-based and near-surface parameters for inland events therefore refer to an extrapolated level. We take this into account when interpreting the regimes, although the largest contrasts between them appear in the 850 and 500 hPa wind speeds (Cliff's $\delta = 0.53$ and 0.46; **Table S4**), and these levels lie above the ground over almost the whole domain (Sect. 3.3, ***L671-676***; Sect. 4, ***L899-903***).

3. The tornado rating in the SINOBAS and ESWD databases were both EF scale or the F scale was used for early reports? If so, how the tornado ratings were homogenised between the two databases?

Response:

Intensity ratings were harmonized into a single set of classes (SD, <EF0 and EF0–EF3) before any analysis, and the procedure is now described in Sect. 2.1 ***(L132-145)***. Almost all SINOBAS reports are rated on the EF scale, whereas ESWD reports carry the scale used by the ESWD at the time of reporting, that is, the original Fujita scale for most of the period and, since 2023, the International Fujita (IF) scale (ESSL, 2023). Ratings were assigned category by category (F0/IF0 to EF0, F1/IF1 to EF1, F2/IF2 to EF2 and F3 to EF3), since the integer IF steps were explicitly designed to be consistent with the categories of the original Fujita scale (ESSL, 2023). The IF scale also introduces half steps (IF0.5, IF1.5), which have no counterpart in the F or EF scales and whose wind-speed ranges span more than one EF class. The three half-step ratings in the catalogue were therefore treated as unrated (SD) rather than assigned to a class by an arbitrary choice. They remain in the occurrence climatology and in the regionalization but, like other unrated events, do not enter the class-specific exceedance calculation. This category-level correspondence is an approximation, because the wind speeds associated with each category differ between scales (Doswell et al., 2009). Its effect on the hazard estimates is nevertheless limited, since rating ambiguities are largest for weak tornadoes (Rodríguez et al., 2026) and the three half-step reports represent less than 1 % of the catalogue (3 of 1132 events).

Citation: https://doi.org/10.5194/egusphere-2026-4053-RC2

References cited only in this response:

- Jackson, D. A.: Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches, Ecology, 74, 2204–2214, https://doi.org/10.2307/1939574, 1993.
- North, G. R., Bell, T. L., Cahalan, R. F., and Moeng, F. J.: Sampling errors in the estimation of empirical orthogonal functions, Mon. Weather Rev., 110, 699–706, https://doi.org/10.1175/1520-0493(1982)110%3C0699:SEITEO%3E2.0.CO;2, 1982.
- Strehl, A. and Ghosh, J.: Cluster ensembles – a knowledge reuse framework for combining multiple partitions, J. Mach. Learn. Res., 3, 583–617, https://www.jmlr.org/papers/v3/strehl02a.html, 2002.
- Wilks, D. S.: Statistical Methods in the Atmospheric Sciences, 4th edn., Elsevier, Amsterdam, https://doi.org/10.1016/C2017-0-03921-6, 2019.