ARTICLE

OPEN

Check for updates

A low-cost post-processing technique improves weather forecasts around the world

Timothy David Hewson ● [1 ✉] & Fatima Maria Pillosu ● [1,2]

Computer-generated weather forecasts divide the Earth's surface into gridboxes, each currently spanning about 400 km^2, and predict one value per gridbox. If weather varies markedly within a gridbox, forecasts for specific sites inevitably fail. Here we present a statistical post-processing method for ensemble forecasts that accounts for the degree of variation within each gridbox, bias on the gridbox scale, and the weather dependence of each. When applying this post-processing, skill improves substantially across the globe; for extreme rainfall, for example, useful forecasts extend 5 days ahead, compared to less than 1 day without post-processing. Skill improvements are attributed to creation of huge calibration datasets by aggregating, globally rather than locally, forecast-observation differences wherever and whenever the observed "weather type" was similar. A strong focus on meteorological understanding also contributes. We suggest that applications for our methodology include improved flash flood warnings, physics-related insights into model weaknesses and global pointwise re-analyses.

[1] ECMWF, Reading, UK. [2] The University of Reading, Reading, UK. ✉email: tim.hewson@ecmwf.int

Weather forecasts nowadays rely heavily on computer-based models, i.e. numerical weather prediction (NWP)[1], and commonly an ensemble of predictions is used, to represent uncertainties[2]. Due to computational power limitations, a gridbox in the best operational global ensembles currently spans about 20 km by 20 km in the horizontal (hereafter: "GM scales" = Global Model scales). So NWP forecasts do not output rainfall (for example) at specific sites, that most customers require, but instead "average rainfall" for much larger gridboxes. This disconnect is an important forecasting problem, which this study addresses. To elaborate, we introduce here the notion of "sub-grid variability", to mean the variation seen amongst all point values observed within the same model gridbox. If sub-grid variability is low then raw NWP forecasts can provide accurate forecasts for points. But if sub-grid variability is high such forecasts inevitably fail.

The most common strategies to address sub-grid variability problems are using a much higher resolution model (e.g. ~2 km*2 km) to minimise them[3], or using calibrated post-processing (PP) techniques to statistically convert from gridbox to point forecasts[4–6]. For predicting rainfall, the parameter central to this article, high-resolution models, whilst showing much more realistic-looking spatial patterns, and exhibiting improvements in forecast skill[7,8], have limited geographical coverage. This is because of computational constraints, which imply that one such model might only cover ~0.2% of the world. For global coverage PP techniques are a better prospect, and they have historically performed well in improving forecasts of dry weather[6,9], but as previous authors themselves acknowledge, many challenges and issues remain (Table 1). Table 1 also highlights how our brand-new approach, described in this study, addresses these points. Ours is a non-local gridbox-analogue approach, formulated via the principles of conditional verification[10], with some structural similarities to quantile regression forests[11,12]. We call the method "ecPoint"—"ec" for ECMWF, i.e. the European Centre for Medium-Range Weather Forecasts, and "Point" for point forecasts.

Sub-grid variability in rainfall is itself very variable (Fig. 1) and relates closely to the weather situation. There are clear-cut physical reasons for this. Dynamics-driven (large-scale) rainfall, often related to atmospheric fronts, arises from steady ascent of moist air across regions typically larger than GM scales (Fig. 1a). As rainfall rates mirror ascent rates, rainfall rate sub-grid variability tends to be small. Conversely instability-driven rainfall (i.e.

showers/convection) arises from localised pockets of rapid ascent, which are typically hundreds of metres to kilometres across. So during convection rainfall rate sub-grid variability, on GM scales, can be very large indeed[13] (Fig. 1b, c).

Rainfall totals arise from integrating rainfall rates over time. Sub-grid variability in totals mostly reduces in proportion to period length. Let us consider, as in many rainfall PP studies, daily to sub-daily time periods (e.g. 6, 12, 24 h), and focus on convection. The intensity, dimensions, density, genesis rate, longevity, and speed of movement of convective cells all impact upon the sub-grid variability in totals. For example, cells moving with speed V, that retain intensity and dimensions for a period t, will deliver stripes in a totals field of length $V*t$. Typical values of V and t might be 15 m s^{-1} and 1 h, giving a stripe 54 km long, which is much greater than GM scales. In such situations we thus get sub-grid variability primarily in one dimension (Fig. 1b). In the limiting case of slow-moving cells, where $V \rightarrow 0$, we retain (large) sub-grid variability in two dimensions (Fig. 1c), and sometimes many locations within "wet gridboxes" stay dry.

The embodiment of ecPoint is that features of the NWP gridbox forecast output (and other global datasets) can tell us what degree of sub-grid variability to expect. For example, NWP output commonly subdivides rainfall into dynamics-driven and convective, and then for convective cases shower movement speed can be approximated by (e.g.) the 700 hPa wind speed. So by using the convective rainfall fraction and the 700 hPa wind speed (two "governing variables") we can distinguish each of the three "gridbox-weather-types" on Fig. 1, to anticipate a priori the expected sub-grid variability, and accordingly convert each forecast for each gridbox into a probabilistic point rainfall prediction (going from red to blue on Fig. 1 PDFs). To our knowledge this general approach, based on first principles of precipitation generation, has not been used before except in a limited way for nowcasting[14] (Table 1, row 11). Another powerful feature of our approach is that each gridbox-weather-type is also associated, via calibration, with a gridscale bias-correction factor.

The logic outlined above could be successfully applied to a single deterministic forecast but in NWP ensembles furnish the most useful predictions[2,15]. So instead we apply separately to each ensemble member, creating an ensemble of probabilistic realisations (or "ensemble of ensembles") that we merge to give the final probabilistic point forecast. Whilst Fig. 1 shows just three gridbox-weather-types, ECMWF's current ecPoint-Rainfall system uses 214 such types, defined in decision tree form.

Table 1 What (previously reported) challenges/issues do classical post-processing methods often face?

	Challenge/issue[a]	Related characteristics of ecPoint
1	Calibration requires ≳20 years of observations[5,6,46,48–51]	Vast training datasets come from just 1 year of data
2	Calibration requires ≳20 years of ensemble re-forecasts[b 6,46,49,50,52]	1 year of re-forecasts, from just one (Control) run, for up to 48 h lead times, is sufficient
3	Lack of climatological stationarity potentially a problem[33,46,51,53,54]	1 year of training is short enough to avoid this
4	Forecasts not possible where no training data is available[46]	Forecasts can be created for all locations
5	Distribution fitting often used may not mirror real data[5,6,12,55]	Nonparametric methods are used
6	Distribution fitting struggles to represent tails well[5,6,51,55,56]	There are no constraints on tail structure
7	Difficult to improve forecasts of extremes[5,6,9,12,52,55–59]	Forecasts of 'extremes' are substantially better
8	Occasional errors in training data may contaminate forecasts[61]	Impacts of such errors are nil or negligible
9	No pointers regarding "reasons" for model errors/biases	"Mapping functions" used denote model characteristics
10	Post-processed spatial output may require smoothing[9,46]	Aesthetically smooth fields arise naturally
11	Large scale/convective precipitation are not disaggregated	An intrinsic and valuable facet of the method
12	Can different modelling systems' biases be compared and addressed?[62,63]	Decision tree reproduction can facilitate this
13	Coverage is not global[c]	Coverage is global

[a]Which issues are relevant depends on the method. Citations denote where an issue has been highlighted, without a clear mechanism for addressing. Studies in which an issue has been highlighted with a mechanism for addressing are as follows: issue no. 6[12], issue 7[60,61], issue 9[60], issue 11[14], issue 13[28].
[b]In the U.S. calibration is the main driving force for running re-forecasts (Tom Hamill, personal communication). ECMWF's re-forecast strategy is different.
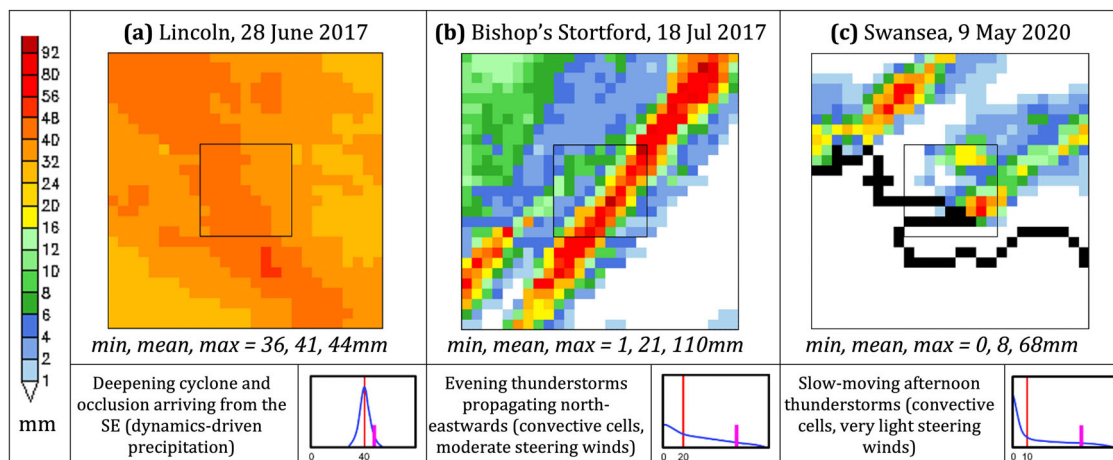[c]Besides the one study[28] no other cited work can have global coverage.

Fig. 1 Three cases of 24 h radar-derived rainfall totals (mm) in the UK illustrating different types of sub-grid variability. a–c Each denote a different case: cells measure 2 × 2 km, black denotes coasts, full frames are 54 × 54 km; legend for 24 h rainfall (mm) applies to all. Central black boxes denote an ECMWF ensemble gridbox (18 × 18 km), for which minimum, mean, and maximum rainfall is shown beneath. Named locations lie approximately mid-panel; all are in regions where relatively flat topography makes radar-derived totals more reliable. Bottom row explains the synoptic situations; inset graphs show, conceptually, how a raw ensemble member forecast (red) for the box should be converted by ecPoint into a probability density function (PDF) for point values (blue) within the box; pink line denotes the 95th percentile; x-scale is linear. Flash floods affected the two regions with red pixel clusters in (**c**)[64]. Radar images are from *netweather.tv*.

Calibration for the post-processing is achieved not by using radar data, but instead rain gauge observations from across the world, in conjunction with short-range Control run re-forecasts, in an innovative, inexpensive, non-local procedure (Table 1, rows 1, 2).

Applications of ecPoint include the many spheres that would benefit from improved probabilistic point forecasts. For rainfall, flash flood prediction is one application, given that we achieve much improved forecasts of localised extremes, as will be shown. ECMWF has been delivering real-time experimental ecPoint-Rainfall products to forecasters since April 2019.

Results

Verification. To have value PP techniques must improve upon raw NWP forecasts. So the performance of 1 year of retrospective forecasts from both the raw NWP and ecPoint-Rainfall systems was compared, using as truth 12 h rainfall observations from both standard SYNOP reports (global coverage) and specialised high-density datasets (certain countries, mainly European[16]). Verification and calibration periods were separate. Although raw model output does not pertain to point values, it is very common to verify using them, as in ECMWF's two headline measures for precipitation[17,18].

Here we utilise categorical verification because threshold setting is common for applications and because forecast products reflect this (see the sections "Case study examples" and "Methods"). In this framework, the two fundamental aspects to assess are reliability (i.e. when x% probability is forecast is the event observed on x% of occasions?) and the capacity to discriminate events. For these, we use respectively, the Reliability component of the Brier Score[19], which is an integral measure of reliability across all issued probabilities, and area under the relative operating characteristic (ROC) curve (ROCA)[15,20].

When verifying probabilistic grid-based forecasts against point measurements one cannot achieve perfect discrimination, because of sub-grid variability. We believe that relative to other discrimination metrics ROCA, also used operationally[18], is much more immune to limitations placed by this. ROCA can exhibit false skill when site climatologies differ[21], so whilst the objective here was to ascertain ecPoint's added value, by *comparing* ROCA scores, a "zero-skill" baseline was also needed, based on local

climatological probabilities (even if these were not available everywhere). Figure 2 displays results for three 12 h accumulation thresholds: 0.2 mm ("dry or not"), 10 mm ("wet"), and 50 mm ("extreme, with flash flood potential").

For both metrics and almost all lead times ecPoint outperforms the raw ensemble. Reliability improvements are particularly striking for the 0.2 mm threshold; this relates to use of a zero multiplier for gridbox totals in convective rainfall situations (e.g. Fig. 1c PDFs). The ecPoint lead-time gains[22] from ROCA, for 0, 10, and 50 mm thresholds are, respectively, about 1, 2, and 8 days (centring on day 5). The particularly large gains for high totals relate, primarily, to weather-type-dependant inclusion of large multipliers for the raw NWP forecast gridbox rainfall (e.g. see blue curve tails on Fig. 1b, c PDFs). To put these results into context, NWP improvements have historically delivered a lead-time gain of about 1 day per decade[1,23]. To have more than "weak potential predictive strength", ROCA must be some way above the zero-skill baseline[24]. So, for 0.2 and 10 mm/12 h thresholds raw ensemble and ecPoint have potential predictive strength out to ~days 7–10. For 50 mm this limit is at least day 5 for ecPoint, whereas the raw ensemble has limited utility even on day 1. ROC curves show that for 10 and 50 mm thresholds the ecPoint system's added value, expressed as ROC area impact, stems from a better handling of the wet tails of the inferred sub-grid point rainfall distributions—e.g. 98th/99th percentiles (see also Supplementary Discussion Section 1 and a more detailed ROC curve presentation in Supplementary Fig. 1).

Equivalent ROCA plots for the tropics only, where site climatologies should be more similar, indicate for point rainfall larger absolute increases in ROCA and larger lead time gains (ROCA climatological skill here is almost unchanged). Nonetheless similar plots for just extratropical regions exhibit lead time gains only ~25% less than the quoted global values. Overall, the slightly better tropical performance arises because for convective weather types, which are more common here, there is more value to be added, which ecPoint does. In improving "convective precipitation" we have addressed the world's number one NWP issue[25].

Reliability, particularly for the raw ensemble, is better for leads >day 1, which may relate to ensemble perturbations being optimised for the medium range. A model spin-up issue also
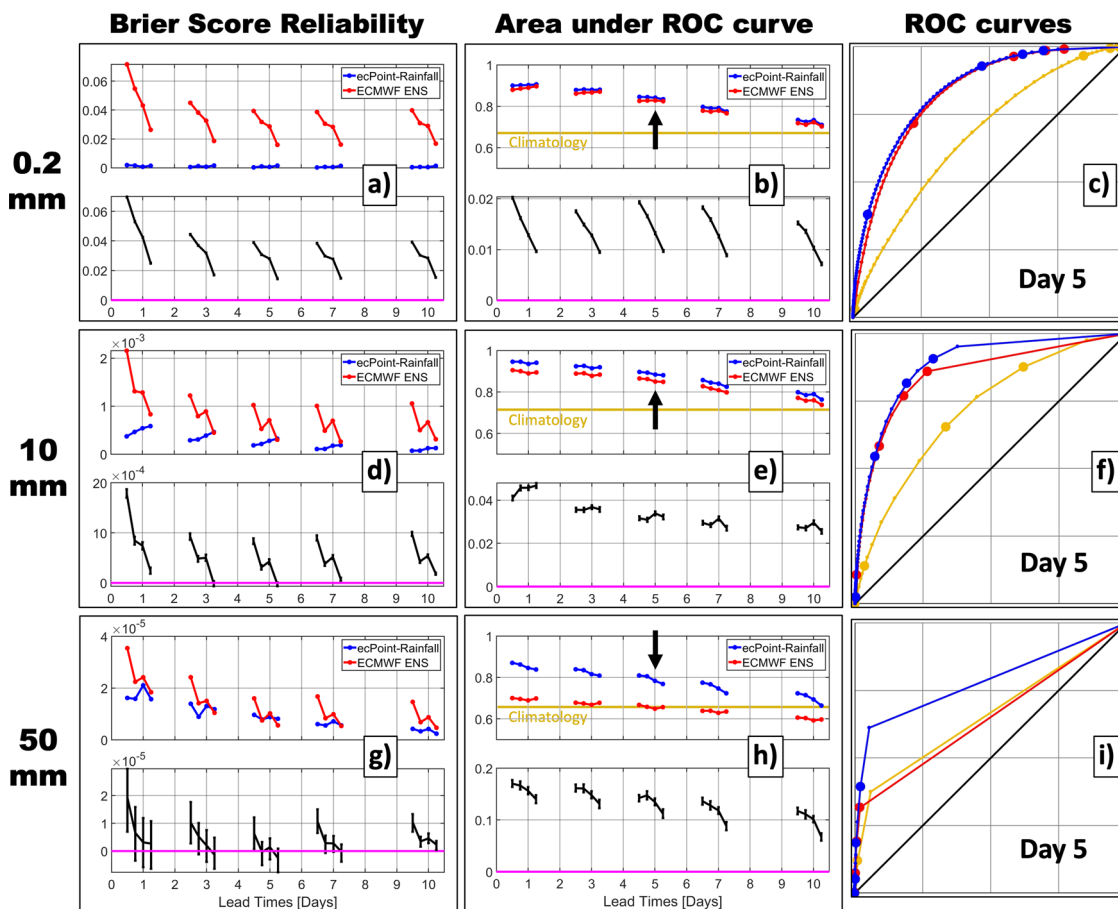
Fig. 2 Category-based verification for 1 year of global gauge observations of 12 h rainfall. a–c, d–f, g–i Signify thresholds of ≥0.2, ≥10, ≥50 mm/12 h, respectively; Red/blue denote raw/point rainfall forecasts, respectively; black profiles beneath denote differences signed such that positive implies point rainfall is better (magenta = 0), with 95% confidence intervals taken from bootstrapping with simple random replacement[65] of daily datasets (1000x). **a**, **d**, **g** show Brier Score reliability component, 0 is optimal; points intersecting vertical gridlines denote 12 h periods ending 00 UTC, others are for end times (left to right) of 12, 18, 06 UTC. **b**, **e**, **h** show area under the ROC curve as a measure of system discrimination ability, larger is better, upper limit is 1; point meaning as for left column; yellow denotes a "baseline", for climatology-based forecasts (for sites where that is available); arrows denote lead time used for right column. **c**, **f**, **i** show day 5 ROC curves (false alarm rate = 0–1 (x) versus hit rate = 0–1 (y)), including climatology; large spots signify probabilities of 2% (topmost), 4%, 10% and 51%. Percentiles 1, 2, …, 99 are used for point rainfall and for climatology.

affects the leftmost points ($T + 0$–12 h). For the 50 mm threshold, although ecPoint looks better overall, the absence of significant reliability gains at most lead times probably relates to information loss in the forecast distribution tail, arising because the largest percentile verified is 99th, even though 99.98th is computed. Although areal warnings of flash flood risk are now issued for relatively low point probabilities, we nonetheless expected little user interest in chances of <1 in 100, and products and verification reflect this. Indeed, ROCA values for ecPoint (and climatology) could both be improved, at low cost, by including higher percentiles, but in practice impact on users would probably be small.

The oscillations seen on all panels on Fig. 2 are a function of UTC time, and relate to irregular observation density coupled with outstanding systematic errors in forecasts of the diurnal cycle of convection[26,27]. Ongoing work with a 6 h accumulation period and a governing variable of local solar time is helping ecPoint to address this deficiency.

In surveying post-processing activities worldwide (see also Supplementary Discussion Section 2) we uncovered just one global method for which a verification comparison with ecPoint was meaningful. That study[28] improved reliability but not resolution for small totals (reference our Fig. 2a, b) and resolution

but not reliability for larger totals (reference our Fig. 2d, e). Verification of very large totals was missing. The study in effect assumes just one weather type. We attribute our much better performance to using multiple types.

Meanwhile, in two 1-year verification comparisons with convection-resolving limited area ensembles, with and without modern post-processing applied, for two relatively mountainous European regions, ecPoint performed as well or better (personal communications: Stephan Hemri, Estibaliz Gascón).

Case study examples. On 25 February 2019 a cyclone (Fig. 3e) delivered extreme rainfall to parts of western Crete: >75 mm/24 h was widely reported, with up to 373 mm/24 h locally (Fig. 3f). Extensive damage occurred, including the collapse of 2 bridges[29]. In the preceding days raw ensemble rainfall forecasts (Fig. 3a) were noisy, jumpy ("Jumpy" is a term regularly used by forecasters: it refers to when forecasts for a given time, produced by consecutive model runs, are inconsistent with one another.) and confusing, with relatively low probabilities of large totals. In relative terms ecPoint (which of course uses the raw ensemble) performs much better (Fig. 3b). First, fields are smoother. Second, probabilities grow over time, are more consistent and are less jumpy. Third, ecPoint's largest probabilities were more focussed
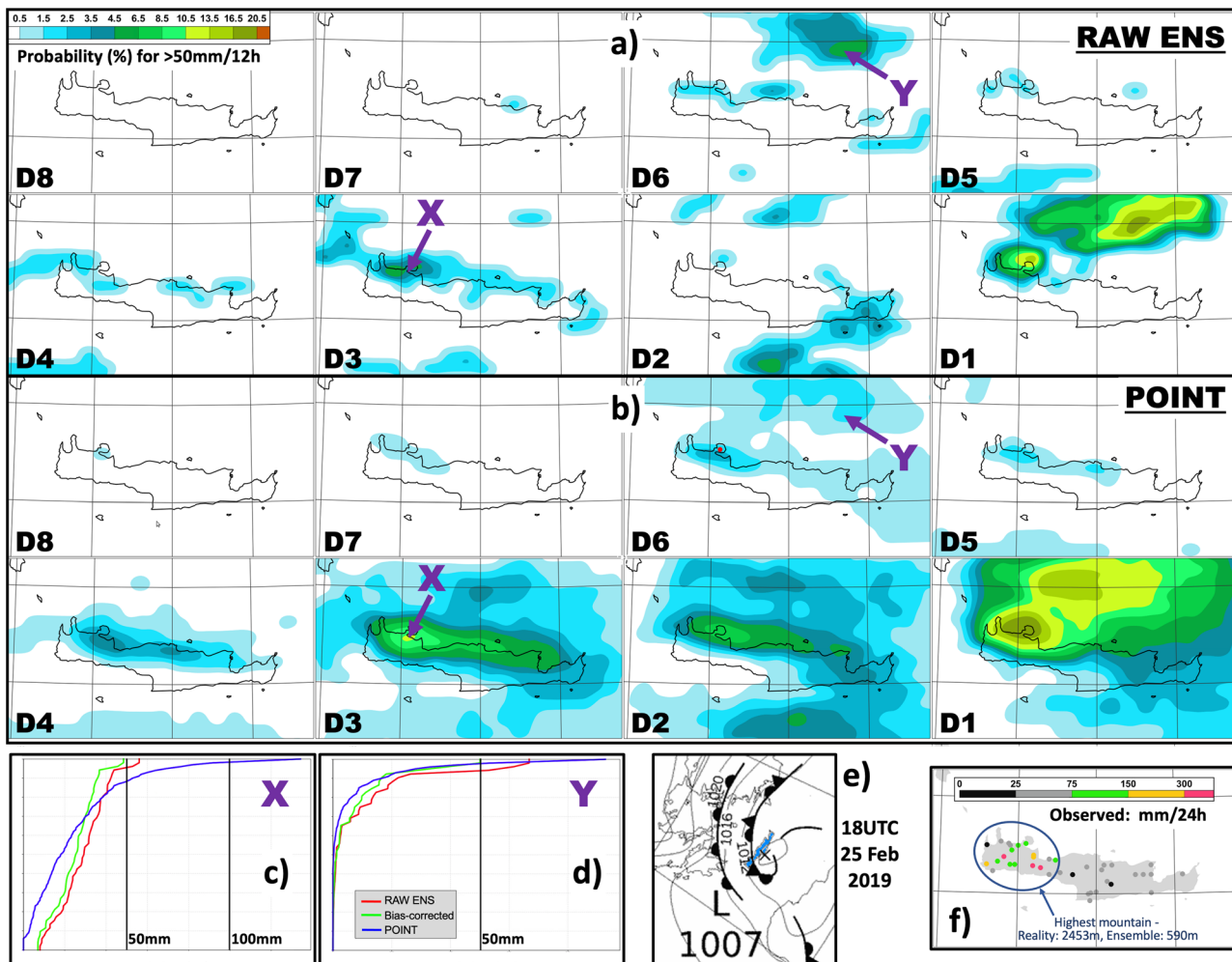
Fig. 3 Evolution of raw ensemble and ecPoint forecasts for a flooding event in Crete. a Forecast probabilities (%), for 12–24 UTC 25 February 2019, for rainfall >50 mm, from raw ensemble; D8, D7, ...D1 (for days 8, 7, ..., 1) denote data times of 00UTC on, respectively, 18, 19 … 25 February. **b** As **a** but for ecPoint. **c** CDFs for 12 h rainfall from the respective D3 forecasts for site X shown on (**a**) and (**b**), for raw ensemble gridbox totals (red), bias-corrected gridbox totals (green) and point rainfall (blue), for 12–24UTC 25th (y-axis spans 0–100% and lies at x = 0 mm). **d** As **c** but for D6 for site Y. **e** UK Met Office surface analysis for 18UTC 25th (Crete in blue). **f** Gauge observations of *24 h rainfall*: 00–24UTC 25 February 2019 (very few 12 h totals are available here).

on western Crete where floods occurred. And fourth, probabilities are higher overall. Conversely, we expect an orographic enhancement shortfall in the raw model (due to lower mountain peaks, Fig. 3f) which was apparently not rectified by ecPoint. Nonetheless, Fig. 3a, b suggest that forecasters could have provided much better warnings here using ecPoint fields, and verification results in Fig. 2 for 50 mm/12 h suggest that this conclusion has general validity.

The probabilities for a large threshold are usually higher in ecPoint output because PP tends to extend the wet tail, at least in (partially) convective situations, as on the cumulative distribution function (CDF) in Fig. 3c. This extension requires, for mathematical reasons, some intra-ensemble consistency. In an alternative case of wet outlier(s), probabilities within the tail can be reduced—e.g. Fig. 3d. This is because gauge totals following convection are most likely to be *less* than the forecast gridbox value, due to the non-Gaussian structure of the expected point total distribution (e.g. see PDFs on Fig. 1b, c).

Figure 1c illustrates smaller-scale flash flood events that were also consistently forecast and localised much better by ecPoint than by the raw ensemble, from 5 days out.

Another feature of ecPoint PP is illustrated in Fig. 4a, b which is for a cyclonic, convective, winter-time case in Norway.

Typically, ecPoint's wetter tail will cross the raw ensemble tail around 85% (e.g. Fig. 3c). But here, even for the 95th percentile, in the north–south chain of maxima (Fig. 4a), ecPoint values are still lower. The reason here is not outliers; it is instead the adjustment for an expected large over-forecasting bias: compare green and red lines in Fig. 4b. Specifically, the main gridbox-weather-type diagnosed here (labelled "44210", see also Supplementary Fig. 2) is itself associated with a substantial gridscale bias-correction. Norwegian forecasters, based on experience, also envisaged that rain in this situation would be over-forecast (personal communication, Vibeke Thyness). Nearby gauge observations (black dots, Fig. 4b), and radar-derived totals seem to support better the ecPoint forecasts.

Figure 4c shows that during calibration the training data for type "44210" came mostly from mountainous regions near coasts, consistent with topographic convective triggering being the rainfall generation mechanism. A likely reason for the over-prediction is that in reality when low-CAPE airmasses impinge on mountains convective cells take time to grow and rain out, whilst in global NWP rainfall is immediate. ecPoint highlights quantitatively the impact. In turn the large biases identified raise important questions about how, for example, related latent heat release overestimates might reduce subsequent broadscale
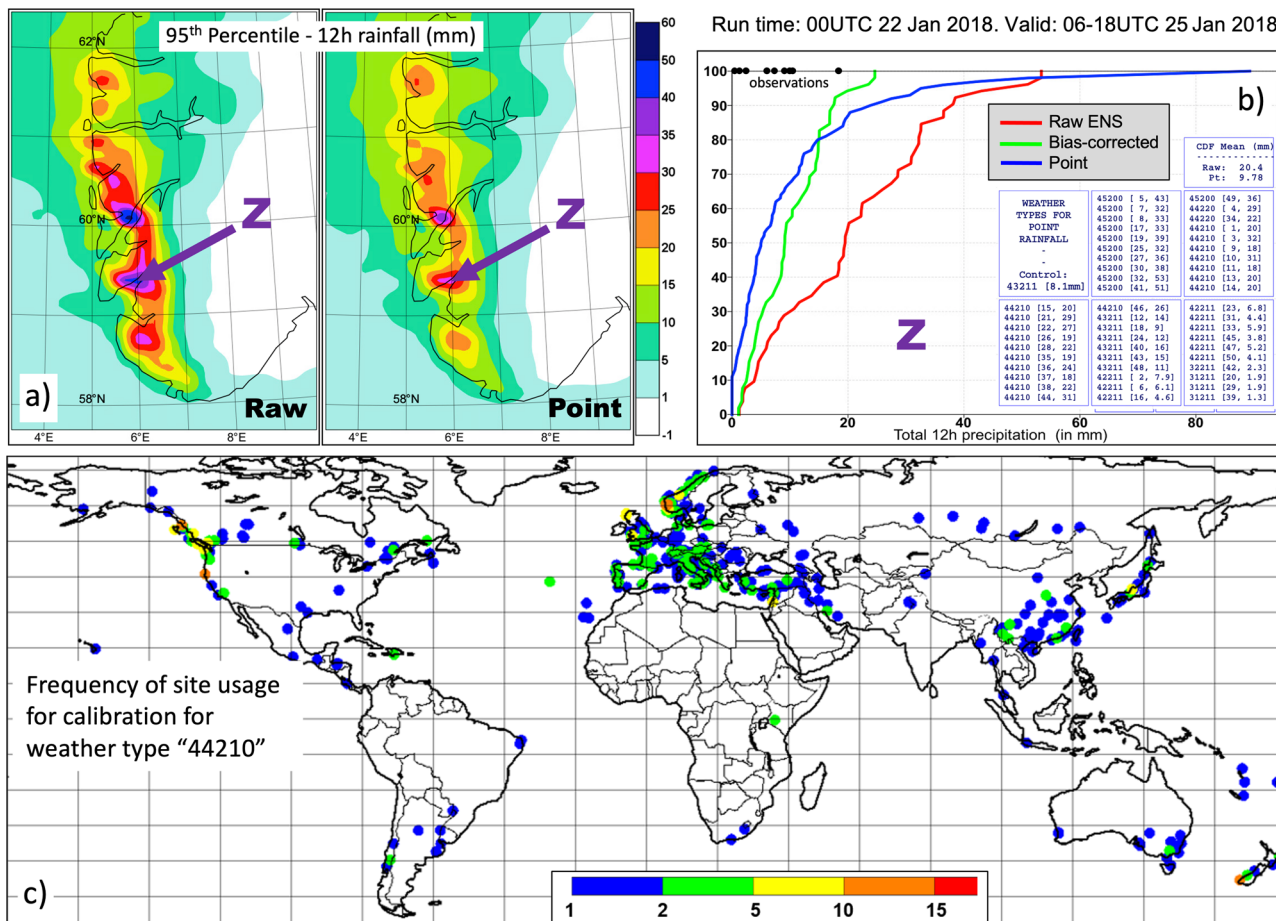
Fig. 4 Day 4 rainfall forecasts for a southwest Norway case, with a calibration site usage map. a 95th percentile of 12 h rainfall (mm) from data time 00UTC 22 January 2018 for 06–18 UTC on 25th, from raw ensemble (left) and ecPoint (right). **b** CDFs of 12 h rainfall for site Z shown on **a**, for same data/valid times as **a**, for raw ensemble gridbox totals (red), bias-corrected gridbox totals (green) and point rainfall (blue); table inset shows: assigned weather type as a 5-digit code, then [member number, forecast rainfall in mm]. **c** Locations of calibration events used to define the gridbox-weather-type "44210".

predictability. The Crete case (Fig. 3) differs because rainfall there had a smaller convective component.

These examples show how ecPoint is rich in physically realistic complexity, which helps deliver much better forecasts on average. Similarly using gridbox-weather-types increases understanding of model behaviour. Other PP methods do not generally have these desirable characteristics.

Figures 3a, b and 4a (and Supplementary Figs. 3a–t and 4a–t) provide examples of the operational ECMWF products. Users can select their own thresholds, compare raw and point forecast fields, and reference user guidelines[30].

Discussion

We describe a completely new PP approach to forecasting weather at sites—ecPoint—to meet customer need in ways that pure global NWP cannot. Our focus has been on rainfall (renowned to be particularly challenging[31,32]) but the philosophy applies also to other variables. Verification results are very positive, even for 50 mm/12 h. Comparable techniques are very scarce. We beat the only real competitor, and match or beat two convection-resolving ensembles with modern post-processing applied. New product characteristics will be appealing to users. Uniquely amongst operational PP methods, ecPoint output covers the world.

One methodological aspect has been pivotal: the use of gridbox-weather-types in lieu of location-specific PP. This delivers, relative to other studies, vast calibration datasets that retain

physical meaning; the many other attractive features of ecPoint stem from this (Table 1). "Weather type" importance has long been recognised, but in the form of country-scale circulation patterns[33,34] which imposes huge constraints on training data size compared to our gridbox-weather-type approach.

Whilst ecPoint can predict local extremes via "remote learning", the lack of *global record* conditions in training data could very occasionally be a constraint. However, using mapping functions means that new global records *can* be predicted whilst in some other PP methods they cannot[12].

For *specific sites* alternative MOS-type techniques (model output statistics) might provide better forecasts than ecPoint, if local topography and/or meteorological characteristics are especially unusual. Similarly, local verification will inevitably reveal some regional differences in ecPoint performance. These should however stimulate future upgrades to the governing variables and the decision tree. More complexity can certainly be included, but in time greater observation coverage will be needed to support this. We therefore plan to exploit more high-density observations, including new crowdsourced data[35], potentially including periods <6 h.

Whilst the simple, expert-driven creation of our decision tree has proved very powerful, there is clearly scope for optimisation. Carefully tailored machine learning tools[36] could be exploited, alongside more complex decision tree branching, and may even become a necessity if governing variable count increases. However, we caution against black-box methods, and enforced statistical complexity, for fear of eroding the capacity to deliver
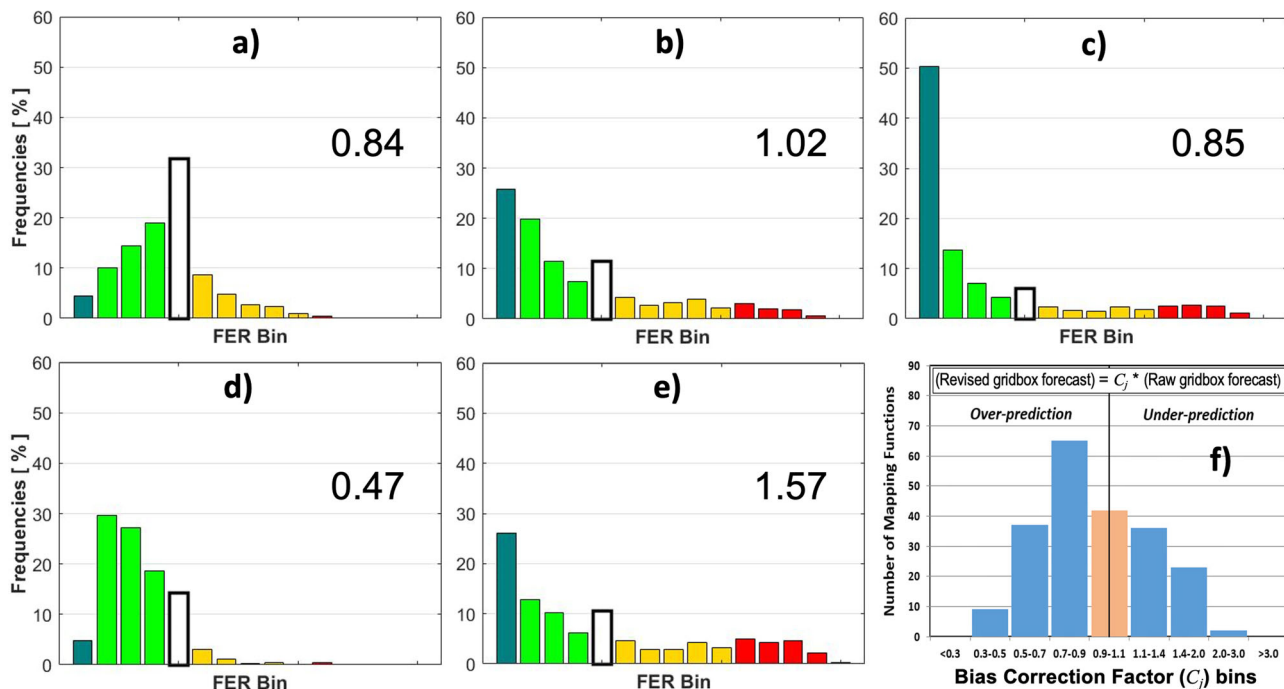
Fig. 5 Mapping function examples, and the bias correction factor distribution. a–e Mapping function examples from the current operational forecast system shown as histograms; large digits show implicit bias correction factors C. Dark green, green, white, yellow, and red bars denote, respectively, FER ranges for "mostly dry" (<−0.99), "over-prediction" (−0.99 to −0.25), "good forecasts" (−0.25 to 0.25), "under-prediction" (0.25–2) and "substantial under-prediction" (2 to ∞). Bar and colour boundaries were subjectively chosen. **a** is for one weather type with $P_{conv} < 0.25$. **b** for $P_{conv} > 0.75$, moderate CAPE (convective available potential energy), strong steering winds, (**c**) for $P_{conv} > 0.75$, moderate CAPE, light winds, (**d**) for $P_{conv} > 0.75$, large forecast totals, moderate winds, low CAPE (=weather type identifier "44210"—see Fig. 4b, c), (**e**) for $0.25 < P_{conv} < 0.5$, modest forecast totals, moderate winds, moderate CAPE. **f** C value distribution for all 214 mapping functions.

meaningful physical insights. It is vital for NWP development to understand the differences between raw and bias-corrected gridbox values. Figure 4b, c provided a clear example of why. Such insights are a major advantage of the ecPoint approach.

ecPoint has many other diverse applications. Its output is currently being blended with post-processed high-resolution limited-area ensemble forecasts[37], to exploit their better representation of some topographic effects (reference Fig. 3f), to deliver better forecasts overall and to seamlessly transition into the medium range. Meanwhile historical probabilistic point rainfall and point temperature re-analyses will be created from the gridscale global re-analysis called ERA5[38], delivering unparalleled representativeness back to 1950. The pointwise climatologies so-derived will themselves have numerous applications, such as providing a reference point (independent of model version/resolution) to see if ecPoint forecasts are *locally* extreme, via the "extreme forecast index" philosophy[39]. Thereby ecPoint forecasts become even more relevant for flash flood prediction. Other applications include inexpensive downscaling of climate projections[40], global tests of hypotheses such as "do cities affect rainfall"[41], and quality-control of observations[42] using a point-CDF-based acceptance window. Improved bias-corrected feeds into hydro-power and hydrological[43] models are also objectives.

So why is ecPoint "low cost"? Providing the control run forecasts we use for calibration is orders of magnitude cheaper than providing the multi-year ensemble re-forecasts needed by many other PP methods. And in operations, ecPoint's own computations are many orders of magnitude cheaper than global NWP alternatives (which in 2021 are very far from being operationally viable anyway). Nonetheless, PP methods would be worthless without high-quality NWP forecasts to provide the input. Indeed, the investment and development needs of global ensembles

remain as strong as ever. The difference now is that by simultaneously exploiting the strong synergistic relationship with ecPoint we can secure much larger forecast improvements for everyone.

Methods

Calibration. In general PP methods need to be calibrated. This requires observations. For ecPoint we use (unadjusted) rain gauge data as this represents points, is commonly used for operational verification[16–18], and has extensive worldwide coverage (over land where rainfall forecast needs are by far the greatest). Whilst issues do exist with gauge data[44] other data types would present far more limitations and complications.

To calibrate we appeal to the concept of conditional verification to create a separate 'mapping function' (**M**) for each of the *m* possible combinations of governing variable ranges (i.e. the *m* gridbox-weather-types). Each such function aims to represent possible outcomes, and indeed a distribution, of point rainfall within a gridbox and so each should represent a different type of sub-grid variability. In our very simple Fig. 1 example there would be three categories (A–C) of "gridbox-weather-type" (or "type") which would have governing variable characteristics like these (P_{conv} = fraction of total precipitation forecast to be convective, V_{700} = 700 hPa wind speed):

- A: $P_{conv} < 0.5$
- B: $P_{conv} > 0.5$ and $V_{700} > 5$ m s^{-1}
- C: $P_{conv} > 0.5$ and $V_{700} < 5$ m s^{-1}

Then mapping functions $\mathbf{M}_A$, $\mathbf{M}_B$, $\mathbf{M}_C$ would represent PDFs of point rainfall within a gridbox for, respectively, types *j* = A–C. To create, in the general case, *m* mapping functions we need to allocate, to one of the *m* types, each and every observation (r_0) taken during a pre-defined period—1 year is sufficient—and over a pre-defined region, ideally the world. At the same time, we must relate these to forecast gridbox rainfall totals (*G*) that inevitably differ, and to do this we introduce a nondimensional metric, the "forecast error ratio" (FER):

$$\text{FER} = (r_0 - G)/G \qquad (\text{requires } G \geq 1 \text{ mm}) \qquad (1)$$

So that $\mathbf{M} = \mathbf{M}(x = \text{FER})$ with $\int_{-1}^{\infty} \mathbf{M} dx = 1$.

Naturally we assign each r_0 value to a gridbox based on location and assign a companion type to that gridbox at that time using values of the selected governing variables. Short-range (unperturbed) control run forecasts provide these values,
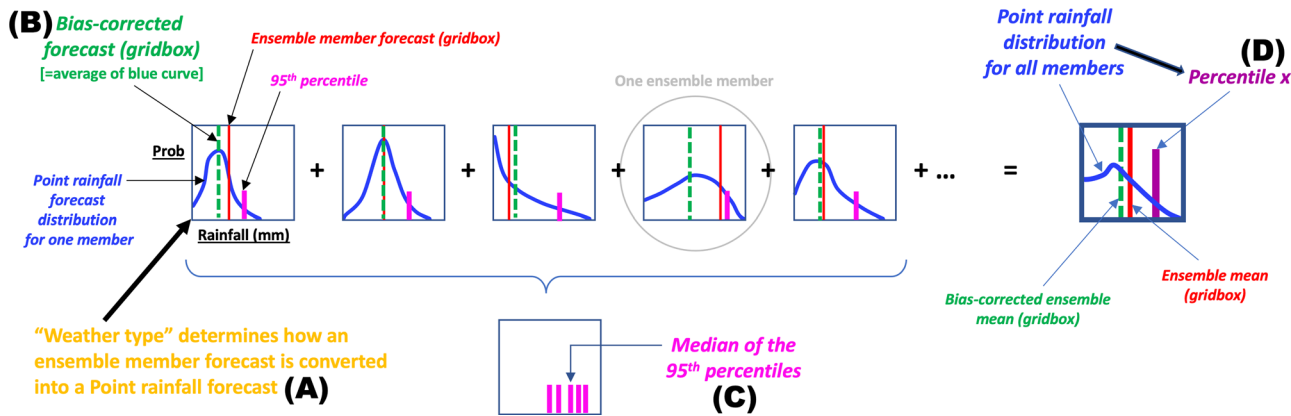
Fig. 6 Schematic showing how an ecPoint (rainfall) forecast is constructed for one gridbox at one time, and outputs that can be created. Blue boxes each show probability density functions (PDFs); bracketed set of 5 represents some ensemble member forecasts, box after "=" shows the final merged "ensemble of ensembles" product. Input gridbox (average) values are shown by red lines and bias-corrected equivalents by dashed green lines; blue curves show ecPoint point PDFs. Pink/purple lines denote percentiles extracted from the respective blue PDFs. Lowermost box shows together the 95th percentiles for each member's point PDF. The primary (single value) ecPoint outputs one can create are labelled in black: (**A**) the weather type identifier index for an ensemble member, (**B**) bias-corrected gridbox rainfall forecast for a member, (**C**) the median rainfall of the 95th (or other) percentiles from all members' point PDFs, (**D**) the full ensemble's point PDF rainfall value at a given percentile x. Output charts for forecasters currently derive from (**D**) with x = 1, 2, …, 99%. The same general principles depicted here also apply for other variables (besides rainfall).

and $G = G_{\text{control}}$. Then, after discarding cases with $G_{\text{control}} < 1$ mm for stability/discretization reasons, each remaining $(G_{\text{control}}, r_0)$ pair furnishes one FER value for the said gridbox-weather-type. By accumulating these we ultimately generate one FER PDF for each mapping function, i.e. $\mathbf{M}_1, \ldots \mathbf{M}_j, \ldots \mathbf{M}_m$. These form the calibration procedure output; five examples are shown in Fig. 5a–e. Equation (1) indicated that NWP "over-prediction" for a rain gauge site was represented by FER < 0 (see dark green and green bars), and "under-prediction" by FER > 0 (see yellow and red bars).

In Fig. 5 examples, sub-grid variability magnitude depends strongly on the gridbox-weather-type. Figure 5a–c, for example, broadly correspond, respectively, to examples in Fig. 1a–c. Variability is lowest in Fig. 5a (distribution roughly Gaussian) and highest, with the greatest likelihood of zeros, in Fig. 5c (distribution roughly exponential). This correspondence supports the use of worldwide gauge observations for our "non-local" calibration. In standard convective situations (Fig. 5b, c) "good forecasts" (white bar) are evidently rare, whilst substantial under-prediction (red), which could lead to "unexpected" flash floods, is relatively common, especially when steering winds are light (Fig. 5c).

Physical reasoning and case studies suggest that *gridscale bias* in raw NWP forecasts can also depend strongly on gridbox-weather-type. As well as predicting sub-grid variability ecPoint also corrects for this bias. The grid-scale bias correction factor (C_j) is implicit in each mapping function; it derives from the "expected value" of the associated FER:

$$C_j = 1 + \int_{-1}^{\infty} x \, \mathbf{M}_j \, dx \qquad (2)$$
$$\approx (\text{mean rainfall observed})/(\text{mean rainfall forecast}) \qquad \text{for type } j.$$

Figure 5f shows the range of C_j values for all mapping functions currently used; for some weather types the ECMWF model seems to markedly over-predict or under-predict rainfall (e.g. Fig. 5d and e, respectively). This is key for ecPoint and is informative also for forecasters and model developers. Although not a definitive inference, Fig. 5f supports the working assumption[27] that *on average* the ECMWF model over-predicts rainfall (although gauge undercatch[44] may be a mitigating factor). The equivalence on line 2 of Eq. (2) assumes $\partial C_j / \partial G \cong 0$. The larger the range of G values the less likely this is to be valid. We reduce ranges by using G as a governing variable for each type j, thereby increasing the efficacy of our model bias interpretation.

A feature of our method is the freedom for the user to select, test, and incorporate any variable that can influence rainfall sub-grid variability and/or bias. The following variable "classes" are included: "*raw model*", "*computed*", "*geographical*", "*astronomical*". The first two relate directly to NWP output, the second two to other datasets. In Supplementary Methods Sections 1–3 these classes are discussed in detail, with examples (see also Supplementary Table 1), and we discuss the assumptions implicit in the approach. Then in Supplementary Methods Sections 4–6 we describe the semi-objective strategy to define governing variable breakpoints and thus decide upon the full set of gridbox-weather-types and the related decision tree (a portion is shown in Supplementary Fig. 2). There is also reference to a new open source graphical user interface (GUI) for performing ecPoint calibration and for building the decision tree.

Unlike classical methods, location is not used in the calibration. This concurs with tests of site separation importance that allocate only 2% weight to this

factor[45]. Removing location facilitates the generation of immense training datasets (matching recommendations[46]) that deliver on average ~10^4 cases for each mapping function. An example will illustrate the powerful implications of this for ecPoint. Consider a Swedish site experiencing a gridbox-weather-type that is a one-in-5-year event there (e.g. including locally extreme convective available potential energy, i.e. CAPE). Such types do occur much more often globally, enough to deliver ~10^4 calibration cases for that type in 1 year of global data. Handling this is computationally straightforward, and moreover the huge case count facilitates our hyper-flexible non-parametric approach. To match this in a simple deterministic MOS approach[4], or a state-of-the-art ensemble approach with 50 supplementary sites[47], would conversely require the impossible—i.e. data for the last 50,000 or 1000 years, respectively.

Forecast production. ecPoint forecast production relies on converting Eq. (1) for FER into a vector form: r_0 becomes $\mathbf{F}_i(r)$, a probabilistic point forecast of rainfall r from member i for the period/gridbox in question, G_i is that member's raw rainfall forecast, and FER becomes $\mathbf{M}_i$ (FER), the mapping function selected for that member. With rearrangement:

$$\mathbf{F}_i(r) = (1 + \mathbf{M}_i(\text{FER}))G_i \qquad \text{where} \qquad \mathbf{M}_i \in \{\mathbf{M}_1, \ldots, \mathbf{M}_m\} \qquad (3)$$

The final probabilistic rainfall forecast vector $\mathbf{F}$, for a point in a given gridbox, is then simply derived using all n ensemble members:

$$\mathbf{F} = \frac{1}{n}\left(\sum_{i=1}^{n} \mathbf{F}_i(r)\right) \qquad (4)$$

The ensemble of ensembles computed operationally currently utilises $m = 214$ mapping functions, with the FER PDF for each simplified into 100 possible outcomes, for computational speed. So for each gridbox/period we arrive at 5100 possible realisations (100*51 members), which are then distilled into percentiles 1,2,…,99 for forecasting. Figure 6 represents the above schematically, whilst Fig. 7 is an example of the geographical distribution of weather types for one ensemble member.

Supplementary Methods Section 7 includes a more detailed discussion of the computational details, and of how the challenges of operational production can be addressed. Then in Supplementary Methods Section 8 we discuss graphical output options and forecaster use of those. Also included there, with verifying data, are examples of how raw ensemble and ecPoint forecast distributions differ, at various lead times in different synoptic settings, over large extratropical and tropical domains (Supplementary Figs. 3 and 4). Supplementary Fig. 5 then demonstrates how significant these distribution differences can be at different leads.

Data availability
Data specific to case Figs. 1c, 3, 4, 7 and verification scores supplementary to Fig. 2 can be found at: 10.5281/zenodo.4728765. ECMWF does not have the legal rights to re-distribute worldwide rainfall observations as used for Fig. 2; some observational data can be obtained from the WMO GTS (World Meteorological Organisation Global Telecommunication System) and from some national archives. Raw model output used in this study is stored in the ECMWF archive (https://www.ecmwf.int/en/forecasts/
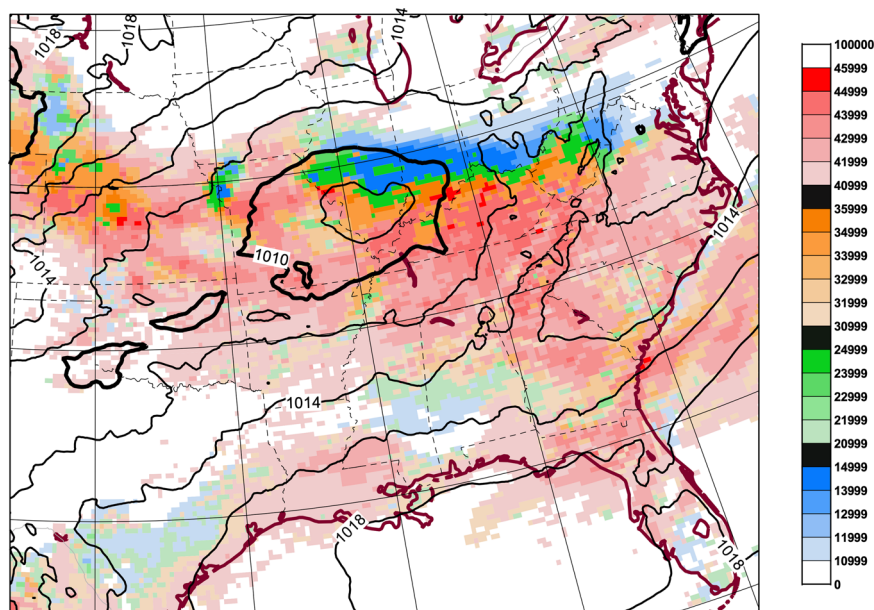
Fig. 7 Example of weather types assigned operationally for a Control run precipitation forecast over the USA for a given 12 h period. Data time is 00UTC 30 July 2020, valid time is 12–24UTC on 30th. Each colour denotes a range of gridbox-weather-type identifiers—see legend and Supplementary Fig. 2 (representing all 214 individually is impractical). Blue, green, orange, red show, respectively, P_{conv} (level 1 on Supplementary Fig. 2) <0.25, 0.25–0.5, 0.5–0.75, 0.75–1; darker shading denotes larger forecast rainfall totals (level 2). Black contours are isobars at 2 hPa intervals from the same Control run, for valid time 18UTC 30th. White cells are dry gridboxes (identifier = "99999"). Note how the main cyclone centre (38°N, 89°W) has large scale rainfall (blue) focussed on its northeastern flank (actually north of a warm front), and predominantly convective rain (red) south of this, with larger forecast rainfall amounts mostly close to and east of the cyclone centre. Each gridbox-weather-type is post-processed using a different mapping function.

datasets/archive-datasets), whilst global ecPoint output files are currently archived separately at ECMWF and are available from T. D.H.

Code availability

Map-based figures were created with "metview" or "metview-python" (https://confluence.ecmwf.int/display/METV). For the post-processing code, calibration files and images of mapping functions see: https://github.com/ecmwf/ecPoint/releases/tag/1.0.0, and for a link to sample data to run this version of the post-processing code see: https://doi.org/10.5281/zenodo.3708501. The calibration software GUI (ecPoint-Calibrate) can be downloaded from https://github.com/esowc/ecPoint-Calibrate, and for sample data to run the software with see: https://doi.org/10.5281/zenodo.4642836.

References

1. Bauer, P., Thorpe, A. & Brunet, G. The quiet revolution of numerical weather prediction. *Nature* https://doi.org/10.1038/nature14956 (2015).
2. Buizza, R. Ensemble forecasting and the need for calibration. In *Statistical Postprocessing of Ensemble Forecasts* (eds Vannitsem, S., Wilks, D. S. & Messner, J. W.) 347 (Elsevier, 2018).
3. Wahl, S. et al. A novel convective-scale regional reanalysis COSMO-REA2: Improving the representation of precipitation. *Meteorol. Z.* https://doi.org/10.1127/metz/2017/0824 (2017).
4. Glahn, H. R. & Lowry, D. A. The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteorol.* **11**, 1203–1211 (1972).
5. Gneiting, T. *Calibration of Medium-range Weather Forecasts*. ECMWF Technical Memoranda 719, 1–28 (ECMWF, 2014)..
6. Flowerdew, J. *Calibration and Combination of Medium-range Ensemble Precipitation Forecasts*. Met Office Forecast. Research Technical Report 567 (Met Office, 2012).
7. Roberts, N. M. & Lean, H. W. Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Mon. Weather Rev* **136**, 78–97 (2008).
8. Mittermaier, M., Roberts, N. & Thompson, S. A. A long-term assessment of precipitation forecast skill using the Fractions Skill Score. *Meteorol. Appl.* **20**, 176–186 (2013).
9. Hamill, T. M. et al. The U.S. National Blend of models for statistical postprocessing of probability of precipitation and deterministic precipitation amount. *Mon. Weather Rev* **145**, 3441–3463 (2017).
10. Ebert, E. et al. Progress and challenges in forecast verification. *Meteorol. Appl.* **20**, 130–139 (2013).
11. Meinshausen, N. Quantile regression forests. *J. Mach. Learn. Res.* **7**, 983–999 (2006).
12. Taillardat, M., Fougeres, A. -L., Naveau, P. & Mestre, O. Forest-based and semi-parametric methods for the postprocessing of rainfall ensemble forecasting. *Weather Forecast.* https://doi.org/10.1175/WAF-D-18-0149.1 (2019).
13. Tapiadorm, F. J. et al. Is precipitation a good metric for model performance. *Bull. Am. Meteorol. Soc.* 223–233 (2019) https://doi.org/10.1175/BAMS-D-17-0218.1
14. Kober, K., Craig, G. C. & Keil, C. Aspects of short-term probabilistic blending in different weather regimes. *Q. J. R. Meteorol. Soc.* https://doi.org/10.1002/qj.2220 (2014).
15. Richardson, D. S. Economic value and skill. In *Forecast Verification: A Practitioner's Guide in Atmospheric Science* (eds Jolliffe, I. T. & Stephenson, D. B.) 165–187 (Wiley, 2003).
16. Haiden, T. & Duffy, S. Use of high-density observations in precipitation verification. *ECMWF Newsl.* https://doi.org/10.21957/hsacrdem (2016).
17. Rodwell, M. J., Richardson, D. S., Hewson, T. D. & Haiden, T. A new equitable score suitable for verifying precipitation in numerical weather prediction. *Q. J. R. Meteorol. Soc.* **136**, 1344–1363 (2010).
18. Haiden, T. et al. *Evaluation of ECMWF Forecasts, Including the 2018 Upgrade*. Technical Memorandum 831 (2018).
19. Toth, Z., Talagrand, O., Candille, G. & Zhu, Y. Probability and Ensemble Forecasts. In *Forecast Verification: A Practitioner's Guide in Atmospheric Science* (eds Ian, T. Jolliffe & David, B. Stephenson) 137–163 (Wiley, 2003).
20. Mason, I. B. Binary events. In *Forecast Verification: a Practitioner's Guide in Atmospheric Science* (eds Jolliffe, I. T. & Stephenson, D. B.) 37–73 (Wiley, 2003).
21. Hamill, T. M. & Juras, J. Measuring forecast skill: is it real skill or is it the varying climatology? *Q. J. R. Meteorol. Soc.* **132**, 2905–2923 (2006).
22. Carroll, E. B. & Hewson, T. D. NWP grid editing at the Met Office. *Weather Forecast* **20**, 1021–1033 (2006).
23. Forbes, R., Haiden, T. & Magnusson, L. Improvements in IFS forecasts of heavy precipitation. *ECMWF Newsl.* 21–26 (2015). https://doi.org/10.21957/jxtonky0
24. Gneiting, T. & Vogel, P. The receiver operating characteristic (ROC) curve. arXiv: 1809.04808 27 (2018).
25. Reynolds, C., Williams, K. & Zadra, A. *WGNE Systematic Error Survey Results Summary*. WMO/WCRP Report (2019).

26. Bechtold, P. et al. Representing equilibrium and nonequilibrium convection in large-scale models. *J. Atmos. Sci.* **71**, 734–753 (2014).
27. Haiden, T. et al. *Use of In Situ Surface Observations at ECMWF*. Technical Memorandum 834 (ECMWF, 2018).
28. Ben Bouallegue, Z., Haiden, T., Weber, N. J., Hamill, T. M. & Richardson, D. S. Accounting for representativeness in the verification of ensemble precipitation forecasts. *Mon. Weather Rev.* **148**, 2049–2062 (2020).
29. Kampouris, N. *One Dead As Floods Cause Extensive Damage Across Crete* (2019) (Greek Reporter, accessed 28 November 2019); https://greece.greekreporter.com/2019/02/26/one-dead-as-floods-cause-extensive-damage-across-crete-video/.
30. Owens, R. G. & Hewson, T. D. *ECMWF Forecast User Guide* (2018). https://doi.org/10.21957/m1cs7h
31. Hemri, S., Scheuerer, M., Pappenberger, F., Bogner, K. & Haiden, T. Trends in the predictive performance of raw ensemble weather forecasts. *Geophys. Res. Lett.* **41**, 9197–9205 (2014).
32. *Guidelines on Ensemble Prediction Systems and Forecasting*. WMO Report 1091 (2012).
33. Van Uytven, E., De Niel, J. & Willems, P. Uncovering the shortcomings of a weather typing method. *Hydrol. Earth Syst. Sci.* **24**, 2671–2686 (2020).
34. Vuillaume, J. F. & Herath, S. Improving global rainfall forecasting with a weather type approach in Japan. *Hydrol. Sci. J.* https://doi.org/10.1080/02626667.2016.1183165 (2017).
35. Muller, C. L. et al. Crowdsourcing for climate and atmospheric sciences: current status and future potential. *Int. J. Climatol.* **35**, 3185–3203 (2015).
36. Haupt, S. E., Pasini, A. & Marzban, C. *Artificial Intelligence Methods in the Environmental Sciences* (Springer, 2009).
37. MISTRAL Consortium. *Definition of MISTRAL Use cases and Services—'Italy Flash Flood'*. 14–16 (2019). http://www.mistralportal.it/wp-content/uploads/2019/06/MISTRAL_D3.1_Use-cases-and-Services.pdf (accessed 27 June 2019)
38. Hersbach, H. et al. The ERA5 global reanalysis. *Q. J. R. Meteorol. Soc.* **146**, 1999–2049 (2020).
39. Zsoter, E. Recent developments in extreme weather forecasting. *ECMWF Newsl.* **107**, 8–17 (2006).
40. Kendon, E. J., Roberts, N. M., Senior, C. A. & Roberts, M. J. Realism of rainfall in a very high-resolution regional climate model. *J. Clim* **25**, 5791–5806 (2012).
41. Han, J.-Y., Baik, J.-J. & Lee, H. Urban impacts on precipitation. *Asia-Pacific J. Atmos. Sci.* **50**, 17–30 (2014).
42. Sciuto, G., Bonaccorso, B., Cancelliere, A. & Rossi, G. Quality control of daily rainfall data with neural networks. *J. Hydrol.* https://doi.org/10.1016/j.jhydrol.2008.10.008 (2009).
43. Verkade, J. S., Brown, J. D., Reggiani, P. & Weerts, A. H. Post-processing ECMWF precipitation and temperature ensemble reforecasts for operational hydrologic forecasting at various spatial scales. *J. Hydrol.* https://doi.org/10.1016/j.jhydrol.2013.07.039 (2013).
44. Sevruk, B. *Methods of Correction For Systematic Error In Point Precipitation Measurement for Operational Use.* WMO Operational Hydrology Report 21 (WMO, 1982).
45. Hamill, T. M., Scheuerer, M. & Bates, G. T. Analog probabilistic precipitation forecasts using GEFS reforecasts and climatology-calibrated precipitation analyses. *Mon. Weather Rev.* **143**, 3300–3309 (2015).
46. Hamill, T. M. Practical aspects of statistical postprocessing. In *Statistical Postprocessing of Ensemble Forecasts* (eds Vannitsem, S., Wilks, D. S. & Messner, J. W.) 347 (Elsevier, 2018).
47. Hamill, T. M. & Scheuerer, M. Probabilistic precipitation forecast postprocessing using quantile mapping and rank-weighted best-member dressing. *Mon. Weather Rev* **146**, 4079–4098 (2018).
48. Marzban, C., Sandgathe, S. & Kalnay, E. MOS, Perfect Prog, and reanalysis. *Mon. Weather Rev* **134**, 657–663 (2006).
49. Wilks, D. S. Statistical methods in the atmospheric sciences: an introduction. *Int. Geophys. Ser.* (1995).
50. Wilks, D. S. & Hamill, T. M. Comparison of ensemble-MOS methods using GFS reforecasts. *Mon. Weather Rev.* **135**, 2379–2390 (2007).
51. Mendoza, P. A. et al. Statistical postprocessing of high-resolution regional climate model output. *Mon. Weather Rev.* **143**, 1533–1553 (2015).
52. Scheuerer, M. & Hamill, T. M. Statistical postprocessing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions. *Mon. Weather Rev.* **143**, 4578–4596 (2015).
53. Wang, Y., Sivandran, G. & Bielicki, J. M. The stationary of two statistical downscaling methods for precipitation under different choices of cross-validation periods. *Int. J. Climatol.* **38**, e330–e348 (2018).
54. Gutiérrez, J. M. et al. Reassessing statistical downscaling techniques for their robust application under climate change conditions. *J. Clim* **26**, 171–188 (2013).
55. Whan, K. & Schmeits, M. Comparing area probability forecasts of (extreme) local precipitation using parametric and machine learning statistical postprocessing methods. *Mon. Weather Rev.* **146**, 3651–3673 (2018).
56. Friedrichs, P., Wahl, S. & Buschow, S. Postprocessing for extreme events. In *Statistical Postprocessing of Ensemble Forecasts* (eds Vannitsem, S., Daniel S. Wilks, D. S. & Messner, J. W.) 347 (Elsevier, 2018).
57. van Straaten, C., Whan, K. & Schmeits, M. Statistical postprocessing and multivariate structuring of high-resolution ensemble precipitation forecasts. *J. Hydrometeorol.* **19**, 1815–1833 (2018).
58. Mylne, K. R., Woolcock, C., Denholm-Price, J. C. W. & Darvell, R. J. Operational calibrated probability forecasts from the ECMWF ensemble prediction system: implementation and verification. In *Joint Session of 16th Conference on Probability and Statistics in the Atmospheric Sciences and of Symposium on Observations, Data Assimilation and Probabilistic Prediction* 113–118 (American Meteorological Society, 2002).
59. Goodwin, P. & Wright, G. The limits of forecasting methods in anticipating rare events. *Technol. Forecast. Soc. Change* **77**, 355–368 (2010).
60. Herman, G. R. & Schumacher, R. S. "Dendrology" in numerical weather prediction: what random forests and logistic regression tell us about forecasting extreme precipitation. *Mon. Weather Rev.* **146**, 1785–1812 (2018).
61. Herman, G. R. & Schumacher, R. S. Money doesn't grow on trees, but forecasts do: forecasting extreme precipitation with random forests. *Mon. Weather Rev.* **146**, 1571–1600 (2018).
62. Zadra, A. et al. Systematic errors in weather and climate models: nature, origins, and ways forward. *Bull. Am. Meteorol. Soc.* **99**, ES67–ES70 (2018).
63. Swinbank, R. et al. The TIGGE project and its achievements. *Bull. Am. Meteorol. Soc.* **97**, 49–67 (2016).
64. Flash floods hit parts of Gorseinon and Carmarthen. *BBC News* (2020). https://www.bbc.co.uk/news/uk-wales-52602177 (accessed 6 August 2020).
65. Wilks, D. S. *Statistical Methods in the Atmospheric Sciences* (Elsevier, 2019).

Author contributions
T.D.H. invented ecPoint, investigated and selected case studies, calculated climatological forecast skill, enacted statistical tests for Supplementary Fig. 5 and prepared the manuscript. F.M.P. investigated governing variable utility, coded up the PP system, performed ecPoint calibration and verification, and contributed to drafting of the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43247-021-00185-9.

Correspondence and requests for materials should be addressed to T.D.H.

Peer review information Communications Earth & Environment thanks the anonymous reviewers for their contribution to the peer review of this work. Primary handling editor: Heike Langenberg

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.