From Drops to Grid: Noise-Aware Spatio-Temporal Neural Process for Rainfall Estimation

Supplementary Material

6. Derivation of ZIG mean and variance

The model outputs a zero-rain probability π_0, from which we define a deterministic per-sample rain indicator

$$p \;=\; \mathbb{1}_{\{1-\pi_0 \geq 0.5\}},$$

so that $p = 1$ denotes predicted nonzero rainfall and $p = 0$ otherwise. Given this indicator, the Zero-Inflated Gamma (ZIG) variable Y is

$$Y \mid (p = 0) = 0, \qquad Y \mid (p = 1) \sim \text{Gamma}(\alpha, \beta),$$

with Gamma mean and variance

$$\mu_\Gamma = \frac{\alpha}{\beta}, \qquad \sigma_\Gamma^2 = \frac{\alpha}{\beta^2}.$$

Because $p \in \{0, 1\}$ is deterministic for each sample, conditioning removes all mixture uncertainty: Y is either identically zero or drawn from a Gamma distribution. Applying the law of total expectation and the law of total variance, respectively:

$$\begin{aligned} \mathbb{E}[Y] &= \mathbb{E}\big[\mathbb{E}[Y \mid p]\big] = (1-p)\cdot 0 + p \cdot \mu_\Gamma = p\,\mu_\Gamma, \\ \mathrm{Var}[Y] &= \mathbb{E}[\mathrm{Var}(Y \mid p)] + \mathrm{Var}(\mathbb{E}[Y \mid p])\,. \end{aligned}$$

The second term vanishes because $\mathbb{E}[Y \mid p]$ is constant for the fixed value of p in a given sample. Therefore,

$$\mathrm{Var}[Y] = p\,\sigma_\Gamma^2.$$

Remark. The more general ZIG variance expression

$$\mathrm{Var}[Y] = p\,\sigma_\Gamma^2 + p(1-p)\mu_\Gamma^2$$

applies when $p \in [0, 1]$ represents a *probabilistic mixture weight*, but the deterministic case with indicator p used here yields the simplified variance $p\,\sigma_\Gamma^2$.

7. Training and dataset details

DropsToGrid employs a U-Net of depth 3 with a kernel size of 3, 32 channels, a bottleneck dropout of 0.1, and transformer blocks of depth 2 with 8 heads of dimension 8. The model contains a total of 192K parameters.

It is trained for up to 50K steps with a batch size of 32, and validation is performed every 1,000 steps. Training uses the AdamW optimizer with a cosine-annealed learning rate starting at 3×10^{-4}, weight decay 1×10^{-4}, betas $(0.9, 0.999)$, and an EMA decay of 0.999. On an NVIDIA H100 PCIe GPU, training requires approximately 3 hours.

The dataset spans January-December 2024 and is divided into training, validation, and test periods of 12, 2, and 2 days, respectively, separated by a 12-hour blackout to prevent temporal leakage. This results in 34,287 training, 911 validation, and 914 test patches.

The target region of $384 \times 384\ \mathrm{km}^2$ is projected onto a 4km/px grid. When multiple stations fall within the same grid cell, the median value is used to mitigate the influence of outliers. Across all samples, up to 902 possible pixels are active in the training set, varying from 797 on January 1st to 884 on December 31st due to station additions, relocations, or temporary outages in the PWS network. The region of interest includes 4,003 land and near-coastal pixels for rainfall estimation. A total of 104 grid cells contain SYNOP stations, with most stations concentrated in urban areas, leading to uneven spatial coverage.

8. Metrics and extended results

We evaluate predictions using Critical Success Index (CSI), Fraction Skill Score (FSS), Frequency Bias Index (FBI), and Continuous Ranked Probability Score (CRPS). Additionally, we report Mean Absolute Error (MAE) and Mean Squared Error (MSE), noting that these are sensitive to the prevalence of no-rain events and may be less informative for highly skewed precipitation distributions [3].

Critical Success Index (CSI) CSI, or Threat Score, measures event detection accuracy by comparing correctly predicted precipitation events to all predicted or observed events. It balances misses and false alarms, making it especially useful for rare-event forecasting such as heavy rainfall. CSI ranges from 0 (no skill) to 1 (perfect prediction).

$$\mathrm{CSI} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}, \tag{1}$$

where TP, FP, and FN are true positives, false positives, and false negatives, respectively, using thresholds of 0.2, 1, 2, 5, and 10 mm/h.

Fraction Skill Score (FSS) FSS evaluates the spatial accuracy of predictions by comparing the fraction of predicted and observed positive events within a local neighborhood, rather than pointwise. It accounts for small spatial displacements, making it suitable for high-resolution precipitation estimation. FSS ranges from 0 (no spatial agreement) to 1 (perfect prediction).

$$\mathrm{FSS} = 1 - \frac{\sum_{i=1}^{H}\sum_{j=1}^{W}(F_{i,j} - O_{i,j})^2}{\sum_{i=1}^{H}\sum_{j=1}^{W}F_{i,j}^2 + \sum_{i=1}^{H}\sum_{j=1}^{W}O_{i,j}^2}, \tag{2}$$

where $F_{i,j}$ and $O_{i,j}$ are the fractions of predicted and observed positives in the neighborhood of pixel (i, j). FSS is computed at thresholds 0.2, 1, 2, 5, and 10 mm/h, and for neighborhood sizes of 2, 10, and 20 pixels.

Frequency Bias Index (FBI) FBI quantifies systematic over- or underprediction of events. Values greater than 1 indicate overprediction, while values below 1 indicate underprediction. A bias of 1 indicates that the predicted frequency matches the observed frequency, though not necessarily the spatial or intensity accuracy.

$$\mathrm{FBI} = \frac{\mathrm{TP} + \mathrm{FP}}{\mathrm{TP} + \mathrm{FN}}, \tag{3}$$

evaluated at the same precipitation thresholds as CSI and FSS (0.2, 1, 2, 5, and 10 mm/h).

Continuous Ranked Probability Score (CRPS) CRPS evaluates the accuracy of a probabilistic forecast by comparing the predicted cumulative distribution function (CDF) to the observed outcome. Lower values indicate better forecast sharpness, calibration, and reliability, rewarding predictions that assign high probability to the correct rainfall intensity. Computation follows closed-form solutions for Gamma distributions [56]. We report CRPS for the probabilistic models (i.e., ablations and deep learning baselines).

For a Zero-Inflated Gamma (ZIG) forecast, which models rainfall R as a mixture of a point mass at zero and a continuous Gamma for positive rainfall, CRPS is computed as:

$$\mathrm{CRPS}_{\mathrm{ZIG}} = \begin{cases} p_{\text{nonzero}} \cdot \mathrm{CRPS}_{\Gamma}(y = 0), & R \leq 0.2 \\ p_{\text{zero}} \cdot R + p_{\text{nonzero}} \cdot \mathrm{CRPS}_{\Gamma}(R), & R > 0.2 \end{cases} \tag{4}$$

where $p_{\text{nonzero}} \in \{0, 1\}$ is the binarized predicted probability of nonzero rainfall, $p_{\text{zero}} = 1 - p_{\text{nonzero}}$, and CRPS_{Γ} is the closed-form CRPS for the Gamma component [56]:

$$\begin{aligned} \mathrm{CRPS}_{\Gamma}(y) = y\,(2F_{\Gamma}(y) - 1) - \frac{\alpha}{\beta}\,(2F_{\Gamma,\alpha+1}(y) - 1) \\ - \frac{\alpha}{\beta\pi}B\Big(\alpha + \frac{1}{2}, \frac{1}{2}\Big) \end{aligned} \tag{5}$$

with F_{Γ} the CDF of the Gamma distribution with parameters (α, β), $F_{\Gamma,\alpha+1}$ the CDF with shape $\alpha + 1$, and $B(\cdot, \cdot)$ the Beta function.

Extended results Tables 4–9 present a more detailed evaluation of DropsToGrid and the gridded baselines against SYNOP stations, including overall metrics, threshold-specific results for CSI, FBI, and FSS, as well as different neighborhood sizes for FSS.

Table 4. Performance comparison across metrics against operational estimators on research-quality SYNOP stations. DropsToGrid achieves superior performance across all metrics. Best results are shown in **bold**, and second-best in *italics*.

	CSI ↑	FSS ↑	FBI ≈ 1	MAE ↓	MSE ↓
OPERA	*0.323*	*0.551*	*0.688*	*0.062*	*0.118*
RainViewer	0.304	0.525	2.231	0.089	0.599
IMERG	0.194	0.425	1.315	0.111	0.261
ERA5	0.203	0.363	0.632	0.084	0.174
DropsToGrid	**0.551 ± 0.006**	**0.795 ± 0.007**	**0.995 ± 0.030**	**0.037 ± 0.001**	**0.070 ± 0.005**

Table 5. Performance comparison across thresholds for CSI against operational estimators on research-quality SYNOP stations. DropsToGrid achieves superior performance across all metrics. Best results are shown in **bold**, and second-best in *italics*.

	CSI ↑				
	0.2 mm/h	1.0 mm/h	2.0 mm/h	5.0 mm/h	10.0 mm/h
OPERA	*0.501*	*0.460*	*0.375*	*0.156*	*0.125*
RadarViewer	0.498	0.446	0.367	0.180	0.029
IMERG	0.280	0.284	0.259	0.102	0.044
ERA5	0.369	0.367	0.281	0.000	0.000
DropsToGrid	**0.674 ± 0.002**	**0.656 ± 0.002**	**0.639 ± 0.007**	**0.444 ± 0.003**	**0.342 ± 0.024**

Table 6. Performance comparison across thresholds for FBI against operational estimators on research-quality SYNOP stations. DropsToGrid achieves superior performance across all metrics. Best results are shown in **bold**, and second-best in *italics*.

	FBI ≈ 1				
	0.2 mm/h	1.0 mm/h	2.0 mm/h	5.0 mm/h	10.0 mm/h
OPERA	*1.093*	0.794	0.657	*0.512*	0.385
RadarViewer	**0.973**	*1.059*	*1.076*	2.122	5.923
IMERG	1.429	1.459	1.364	1.517	**0.808**
ERA5	1.346	1.081	0.731	0.000	0.000
DropsToGrid	0.887 ± 0.005	**1.054 ± 0.005**	**1.041 ± 0.020**	**1.238 ± 0.038**	*0.756 ± 0.101*

Table 7. Performance comparison across thresholds for FSS (neighborhood size: 2px) against operational estimators on research-quality SYNOP stations. DropsToGrid achieves superior performance across all metrics. Best results are shown in **bold**, and second-best in *italics*.

	FSS (size: 2px) ↑				
	0.2 mm/h	1.0 mm/h	2.0 mm/h	5.0 mm/h	10.0 mm/h
OPERA	*0.671*	*0.635*	*0.550*	0.266	*0.220*
RadarViewer	0.669	0.623	0.543	*0.310*	0.056
IMERG	0.440	0.444	0.415	0.184	0.084
ERA5	0.542	0.539	0.443	0.000	0.000
DropsToGrid	**0.809 ± 0.001**	**0.796 ± 0.002**	**0.784 ± 0.005**	**0.622 ± 0.003**	**0.510 ± 0.031**

Table 8. Performance comparison across thresholds for FSS (neighborhood size: 10px) against operational estimators on research-quality SYNOP stations. DropsToGrid achieves superior performance across all metrics. Best results are shown in **bold**, and second-best in *italics*.

	FSS (size: 10px) ↑				
	0.2 mm/h	1.0 mm/h	2.0 mm/h	5.0 mm/h	10.0 mm/h
OPERA	*0.780*	*0.756*	0.679	0.338	*0.261*
RadarViewer	0.777	0.752	*0.680*	*0.419*	0.055
IMERG	0.549	0.564	0.535	0.284	0.214
ERA5	0.628	0.640	0.557	0.000	0.000
DropsToGrid	**0.900 ± 0.001**	**0.897 ± 0.001**	**0.892 ± 0.004**	**0.782 ± 0.001**	**0.597 ± 0.039**

Table 9. Performance comparison across thresholds for FSS (neighborhood size: 20px) against operational estimators on research-quality SYNOP stations. DropsToGrid achieves superior performance across all metrics. Best results are shown in **bold**, and second-best in *italics*.

	FSS (size: 20px) ↑				
	0.2 mm/h	1.0 mm/h	2.0 mm/h	5.0 mm/h	10.0 mm/h
OPERA	*0.843*	0.816	0.740	0.418	0.289
RadarViewer	0.839	*0.829*	*0.757*	*0.488*	0.076
IMERG	0.628	0.662	0.636	0.385	*0.359*
ERA5	0.700	0.730	0.664	0.000	0.000
DropsToGrid	**0.943 ± 0.000**	**0.946 ± 0.001**	**0.943 ± 0.003**	**0.860 ± 0.003**	**0.649 ± 0.040**

9. Baseline gridded products

For evaluation, we use operational and reanalysis gridded rainfall products as reference baselines rather than ground truth. To ensure comparability, all baselines are resampled to a uniform 4 km grid using bilinear interpolation and converted to hourly rainfall accumulations (mm). Further details on each gridded baseline product are provided below.

OPERA Odyssey rainfall accumulation. The EUMETNET OPERA program [29, 53] provides pan-European radar composites combining data from over 160 national radars. The 1-hour accumulation product has 2 km resolution and 15-minute updates. Rain rate is derived from reflectivity using the Marshall-Palmer $Z = aR^b$ relation ($a = 200$, $b = 1.6$) [40]. Quality filtering employs anomaly removal [47], clutter [57], and beam-blockage [21] corrections. Maximum accumulation is set to 100 mm.

RainViewer reflectivity. RainViewer [3] provides near-real time radar composites every 10 minutes at 2km resolution, aggregating over 1,000 stations worldwide. Reflectivity (dBZ) is clipped to [-1, 64] and converted to intensity (mm/h) via Marshall-Palmer [40]. Hourly accumulations are computed by averaging the six 10-min intensity maps.

IMERG. The GPM IMERG product [28] merges passive microwave, infrared, and radar data from multiple satellites to provide global rainfall estimates at 0.1° and 30-min intervals. We use the Final Run V07 dataset, bias-corrected to match rain-gauge climatologies. Hourly accumulations are obtained by averaging the two 30-min frames per hour.

ERA5. ERA5 [24, 61] is ECMWF's global reanalysis, combining a frozen numerical model with a consistent data assimilation scheme. Although it does not directly assimilate European rainfall observations, estimates are derived from physically consistent atmospheric fields (humidity, geopotential, winds). ERA5 provides total rainfall, including convective, large-scale, and evaporative effects, at 0.25° resolution and hourly frequency.

Climate (DMI gridded product). The DMI *Climate* dataset [55] provides hourly gridded rainfall at 10 km resolution. Fields are generated using Inverse Distance Weighting (IDW) interpolation of quality-controlled gauges, followed by Gaussian smoothing and a coastal-inland correction to adjust distance metrics across land-sea boundaries. The gauge network used to generate the gridded product includes the SYNOP stations previously mentioned.

[3] https://www.rainviewer.com

10. Visualizations

Figures 5-8 show visual comparisons of rainfall estimates from DropsToGrid and several operational baselines. DropsToGrid is derived from crowd-sourced PWS stations and RainViewer radar.

The baselines include OPERA radar accumulations, RainViewer reflectivity estimates, IMERG satellite retrievals, ERA5 reanalysis, and DMI's gridded *Climate* product. CSI is reported for all products against SYNOP stations, except for Climate, which is excluded since it incorporates the evaluation SYNOP stations in its gridding.

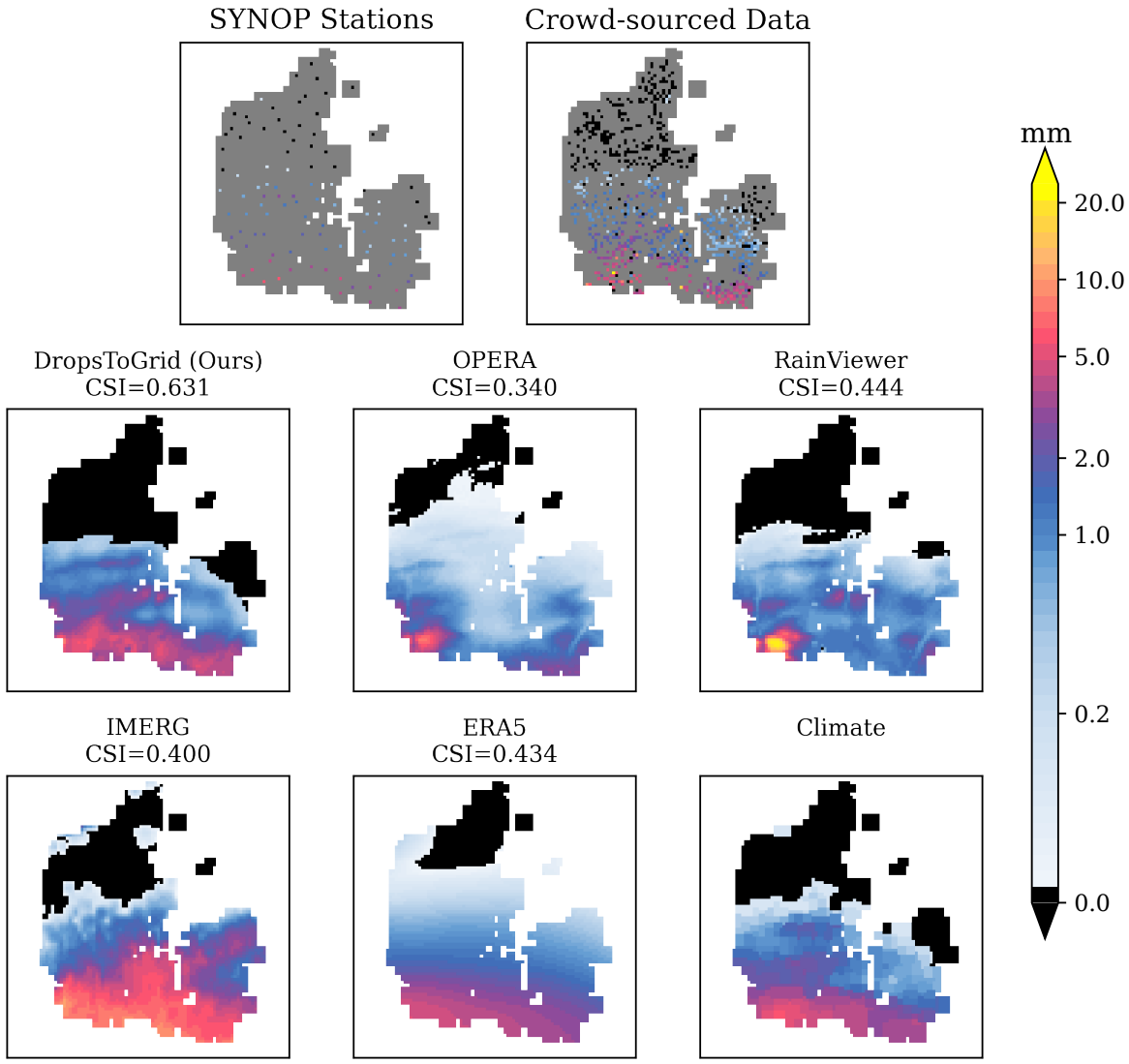

Figure 5. Comparison of rainfall estimators against research-quality SYNOP stations.

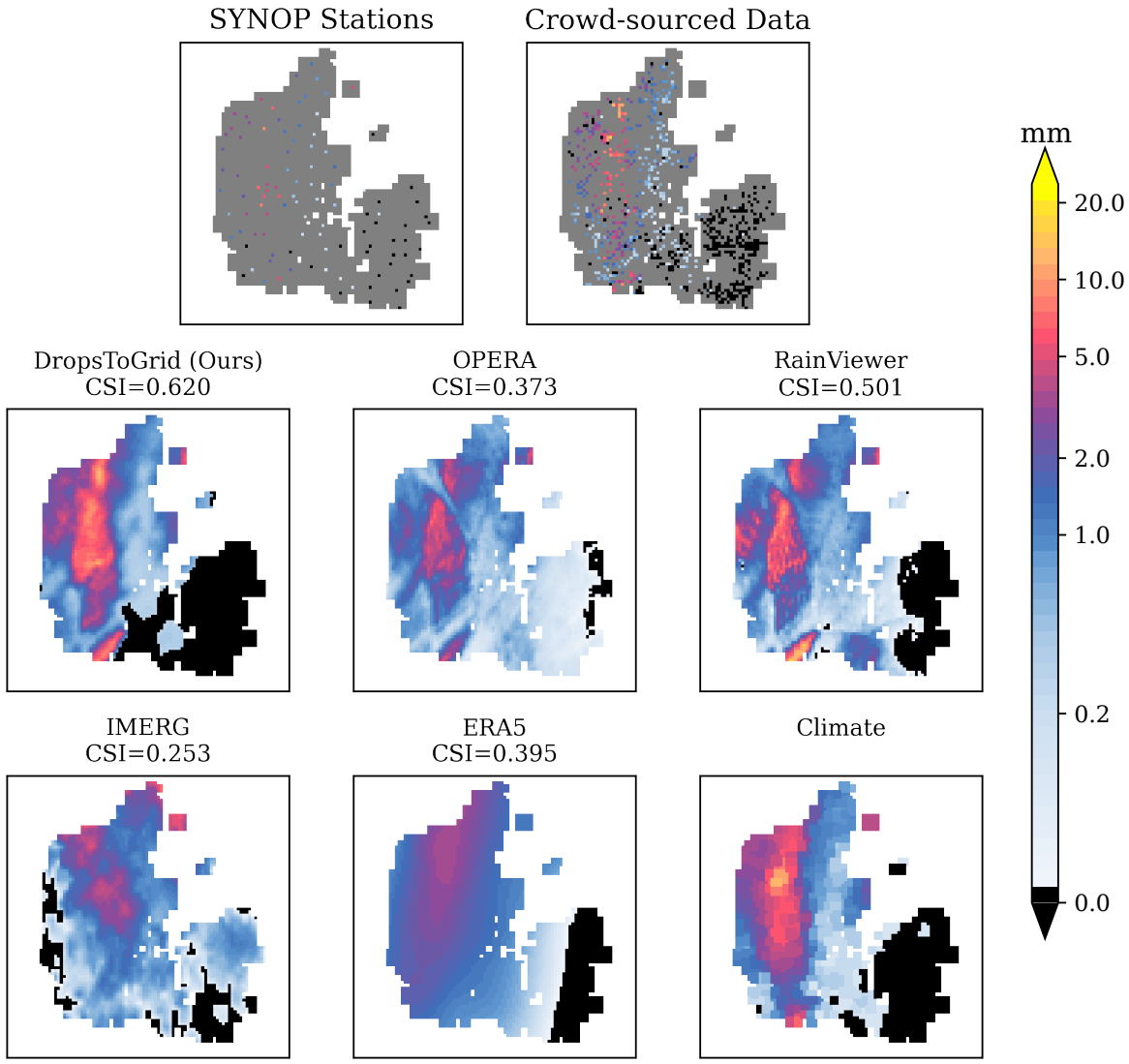

Figure 6. Comparison of rainfall estimators against research-quality SYNOP stations.

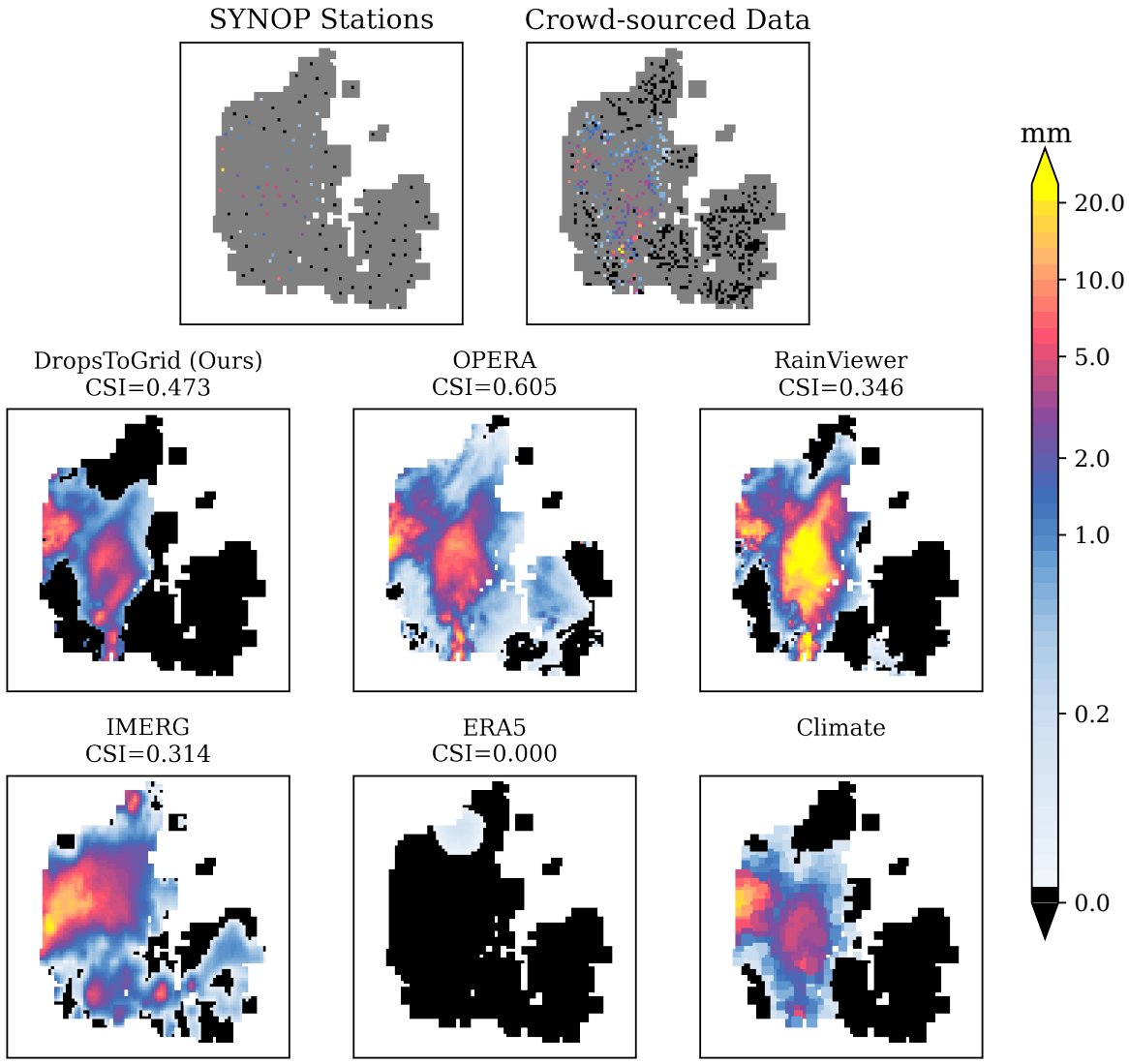

Figure 7. Comparison of rainfall estimators against research-quality SYNOP stations.

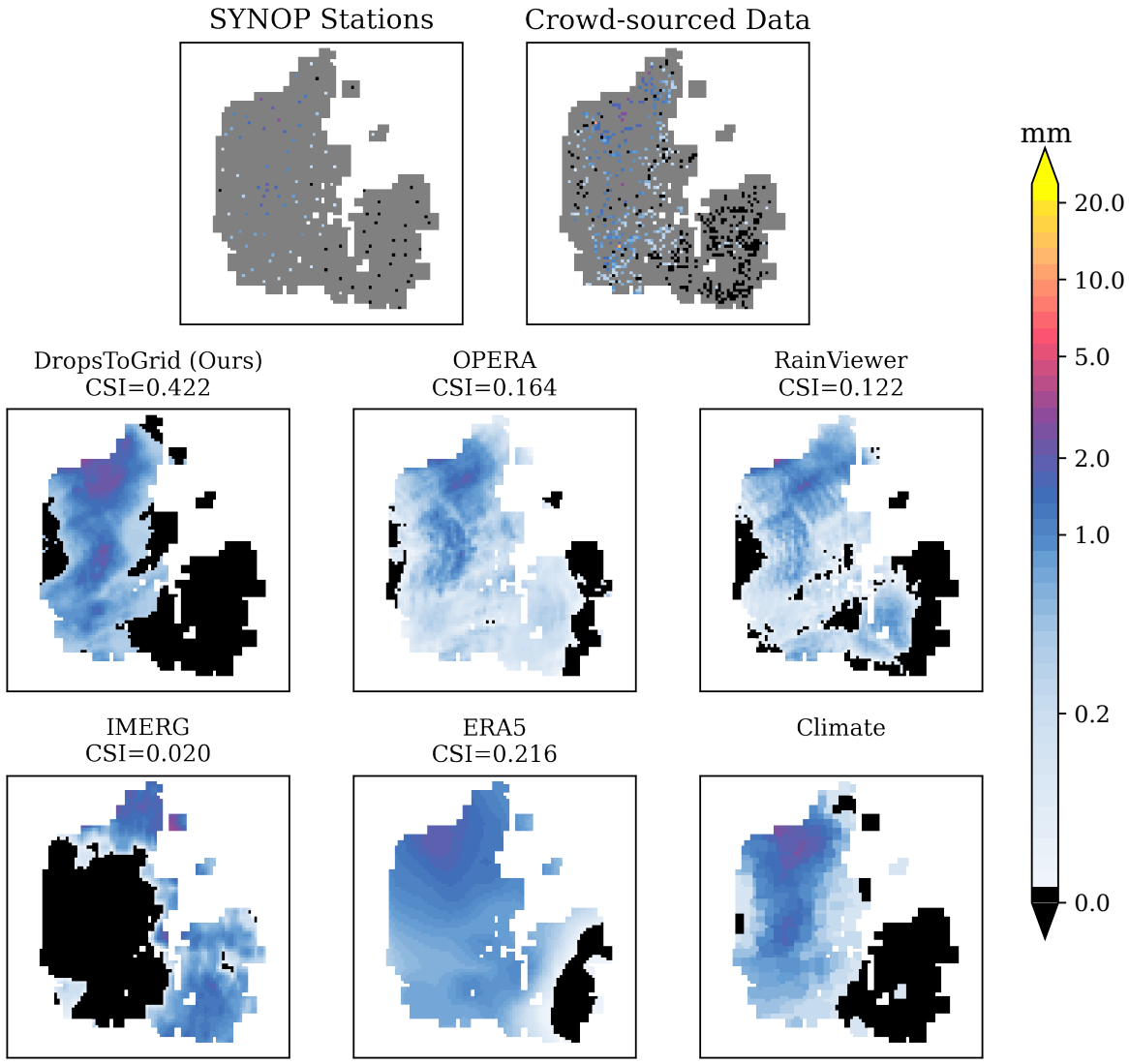

Figure 8. Comparison of rainfall estimators against research-quality SYNOP stations.

11. Deep Learning baselines

Beyond the ConvCNP and SwinTNP variants used in the main paper, we evaluate additional baselines. The MM setting uses only station history as input (no radar), while the OOTG setting uses radar and current-time station readings (no history). We further include the translation-equivariant SwinTNP (SwinTNP_TE) and the approximately translation-equivariant version (SwinTNP_ATE) from Gridded-TNP [6] in both settings. Finally, we add an extended ALL ConvCNP model that uses both radar and station history to provide a closer comparison to DropsToGrid in terms of data sources.

Across both the PWS holdout and SYNOP stations (Tables 10–11), DropsToGrid achieves the strongest results on almost all metrics, with higher spatial skill (CSI, FSS), better calibration (CRPS), and lower errors (MAE, MSE). ConvCNP models generally outperform the SwinTNP variants in both MM and OOTG settings, likely because they better capture the local structure of rainfall. Even when provided with both radar and station history, the ALL ConvCNP baseline remains below DropsToGrid, underscoring the benefits of DropsToGrid's processing and fusion design.

Table 10. Performance comparison across metrics against deep learning baselines on PWS holdout stations. DropsToGrid achieves superior performance across most metrics. Best results are shown in **bold**, and second-best in *italics*.

	CSI ↑	FSS ↑	CRPS ↓	FBI ≈ 1	MAE ↓	MSE ↓
MM						
ConvCNP	0.484 ± 0.012	0.769 ± 0.035	0.029 ± 0.000	0.849 ± 0.061	0.038 ± 0.001	0.124 ± 0.003
SwinTNP	0.407 ± 0.002	0.676 ± 0.003	0.034 ± 0.000	0.766 ± 0.029	0.046 ± 0.001	0.145 ± 0.002
SwinTNP_TE	0.412 ± 0.010	0.672 ± 0.021	0.034 ± 0.000	0.814 ± 0.071	0.047 ± 0.001	0.147 ± 0.002
SwinTNP_ATE	0.418 ± 0.014	0.700 ± 0.017	0.034 ± 0.001	0.772 ± 0.039	0.045 ± 0.000	0.145 ± 0.003
OOTG						
ConvCNP	*0.486 ± 0.007*	0.748 ± 0.013	*0.027 ± 0.000*	0.752 ± 0.015	**0.035 ± 0.000**	*0.115 ± 0.001*
SwinTNP	0.433 ± 0.001	0.711 ± 0.006	0.032 ± 0.000	0.805 ± 0.020	0.044 ± 0.000	0.140 ± 0.002
SwinTNP_TE	0.428 ± 0.002	0.684 ± 0.010	0.032 ± 0.000	0.763 ± 0.012	0.043 ± 0.000	0.141 ± 0.002
SwinTNP_ATE	0.434 ± 0.004	0.696 ± 0.010	0.032 ± 0.000	0.798 ± 0.045	0.043 ± 0.001	0.142 ± 0.003
ALL						
ConvCNP	0.485 ± 0.022	*0.775 ± 0.035*	0.029 ± 0.000	**0.879 ± 0.109**	0.039 ± 0.001	0.146 ± 0.022
DropsToGrid	**0.532 ± 0.002**	**0.819 ± 0.000**	**0.026 ± 0.000**	*0.877 ± 0.019*	*0.035 ± 0.000*	**0.112 ± 0.001**

Table 11. Performance comparison across metrics against deep learning baselines on research-quality SYNOP stations. DropsToGrid achieves superior performance across most metrics. Best results are shown in **bold**, and second-best in *italics*.

	CSI ↑	FSS ↑	CRPS ↓	FBI ≈ 1	MAE ↓	MSE ↓
MM						
ConvCNP	0.507 ± 0.012	*0.760 ± 0.020*	0.025 ± 0.000	0.863 ± 0.050	0.038 ± 0.001	0.071 ± 0.001
SwinTNP	0.424 ± 0.005	0.666 ± 0.008	0.030 ± 0.000	0.797 ± 0.032	0.045 ± 0.001	0.093 ± 0.002
SwinTNP_TE	0.423 ± 0.014	0.653 ± 0.027	0.030 ± 0.000	0.834 ± 0.089	0.046 ± 0.001	0.094 ± 0.001
SwinTNP_ATE	0.437 ± 0.010	0.679 ± 0.013	0.030 ± 0.001	0.780 ± 0.037	0.045 ± 0.000	0.091 ± 0.003
OOTG						
ConvCNP	*0.522 ± 0.015*	0.754 ± 0.021	*0.023 ± 0.000*	0.823 ± 0.031	**0.035 ± 0.000**	**0.067 ± 0.000**
SwinTNP	0.457 ± 0.005	0.700 ± 0.007	0.028 ± 0.000	0.859 ± 0.035	0.043 ± 0.001	0.087 ± 0.001
SwinTNP_TE	0.439 ± 0.001	0.655 ± 0.005	0.028 ± 0.000	0.790 ± 0.009	0.042 ± 0.000	0.087 ± 0.001
SwinTNP_ATE	0.439 ± 0.007	0.665 ± 0.014	0.028 ± 0.000	0.824 ± 0.045	0.043 ± 0.001	0.087 ± 0.003
ALL						
ConvCNP	0.513 ± 0.007	0.757 ± 0.006	0.025 ± 0.000	*0.987 ± 0.144*	0.039 ± 0.001	0.090 ± 0.016
DropsToGrid	**0.551 ± 0.006**	**0.795 ± 0.007**	**0.023 ± 0.000**	**0.995 ± 0.030**	*0.037 ± 0.001*	*0.070 ± 0.005*

12. Ablation

Tables 12 and 13 summarize the ablation studies conducted to assess the contribution of each component of DropsToGrid on the PWS holdout stations and the research-grade SYNOP stations, respectively. The two primary ablations discussed in the main paper examine (i) replacing the carefully designed fusion bottleneck with a standard convolution that simply stacks all latent source representations in the input (*no_bottleneck*), and (ii) allowing input stations to also serve as target stations during training (*target_inputs*), in contrast to DropsToGrid's strategy of excluding inputs from the prediction targets to avoid direct input-output mapping and mitigate noisy observations. We additionally evaluate variants without station history (*no_stations*), without radar (*no_radar*), and with a transformer using standard (non–translation-equivariant) attention (*no_te*). Regarding the output distribution, beyond the *target_inputs* ablation, we test a plain gamma distribution that omits the zero-inflation component (*gamma*), as well as a Gaussian output distribution (*gaussian*), to highlight the importance of modeling the highly skewed, zero-dominated nature of rainfall.

As noted in the main paper, the specialized fusion bottleneck is critical, with *no_bottleneck* showing consistent degradation across all metrics. Although the effects are smaller in magnitude, radar information, station history, and translation-equivariant attention each provide measurable gains, from pixelwise skill (CSI) to calibration of the distribution (CRPS) and overall error (MAE/MSE). Preventing direct input–output mapping is also important, particularly given the noise in PWS observations, as reflected in the significant performance drop against SYNOP stations in *target_inputs*. Finally, the *gamma* and *gaussian* variants confirm that an appropriate output distribution is essential for rainfall, whose statistics are highly skewed and dominated by zero-rain cases.

Table 12. Ablation experiments on PWS holdout stations. DropsToGrid achieves superior performance across most metrics. Best results are shown in **bold**, and second-best in *italics*.

	CSI ↑	FSS ↑	CRPS ↓	FBI ≈ 1	MAE ↓	MSE ↓
DropsToGrid	**0.532 ± 0.002**	0.819 ± 0.000	**0.026 ± 0.000**	0.877 ± 0.019	**0.035 ± 0.000**	**0.112 ± 0.001**
no_bottleneck	0.520 ± 0.015	0.804 ± 0.025	0.027 ± 0.000	0.868 ± 0.042	0.036 ± 0.001	0.115 ± 0.005
no_stations	*0.530 ± 0.004*	0.825 ± 0.006	*0.026 ± 0.000*	0.905 ± 0.006	0.036 ± 0.000	0.113 ± 0.003
no_radar	0.530 ± 0.004	**0.829 ± 0.004**	0.027 ± 0.000	*0.905 ± 0.011*	0.036 ± 0.000	0.115 ± 0.001
no_te	0.530 ± 0.005	*0.826 ± 0.005*	0.026 ± 0.000	0.871 ± 0.018	*0.036 ± 0.000*	*0.113 ± 0.001*
target_inputs	0.514 ± 0.008	0.816 ± 0.006	0.028 ± 0.001	**0.996 ± 0.044**	0.039 ± 0.001	0.149 ± 0.007
gamma	0.471 ± 0.016	0.727 ± 0.024	0.032 ± 0.000	0.705 ± 0.052	0.038 ± 0.001	0.118 ± 0.003
gaussian	0.414 ± 0.006	0.637 ± 0.016	0.045 ± 0.003	0.635 ± 0.027	0.041 ± 0.001	0.129 ± 0.001

Table 13. Ablation experiments on research-quality SYNOP stations. DropsToGrid achieves superior performance across most metrics. Best results are shown in **bold**, and second-best in *italics*.

	CSI ↑	FSS ↑	CRPS ↓	FBI ≈ 1	MAE ↓	MSE ↓
DropsToGrid	**0.551 ± 0.006**	*0.795 ± 0.007*	**0.023 ± 0.000**	**0.995 ± 0.030**	*0.037 ± 0.001*	**0.070 ± 0.005**
no_bottleneck	0.529 ± 0.023	0.770 ± 0.023	*0.024 ± 0.000*	0.966 ± 0.058	**0.037 ± 0.000**	0.077 ± 0.006
no_stations	0.549 ± 0.003	0.794 ± 0.004	0.024 ± 0.000	1.032 ± 0.019	0.037 ± 0.000	0.078 ± 0.001
no_radar	*0.551 ± 0.009*	**0.796 ± 0.009**	0.024 ± 0.000	1.026 ± 0.025	0.037 ± 0.000	0.072 ± 0.002
no_te	0.542 ± 0.002	0.789 ± 0.004	0.024 ± 0.000	*0.989 ± 0.006*	0.037 ± 0.000	*0.072 ± 0.003*
target_inputs	0.505 ± 0.009	0.755 ± 0.008	0.029 ± 0.001	1.198 ± 0.044	0.042 ± 0.001	0.119 ± 0.014
gamma	0.485 ± 0.014	0.719 ± 0.022	0.033 ± 0.000	0.769 ± 0.041	0.038 ± 0.000	0.075 ± 0.003
gaussian	0.409 ± 0.009	0.600 ± 0.016	0.046 ± 0.002	0.622 ± 0.028	0.041 ± 0.001	0.081 ± 0.002

13. Station analysis

To assess how varying observational coverage affects DropsToGrid, we study performance under progressively reduced densities of input PWS stations. Starting from all 902 pixels with PWS data, we randomly mask stations in 10% increments, using nested masks so that each higher-density configuration contains all stations from the previous one. We repeat the experiment with three seeds and average the results, where each seed influences the station masking and model. For each density level, we evaluate the complete test set against SYNOP stations using CRPS and CSI with only the corresponding subset of stations as input.

DropsToGrid performs best when all stations are available, but remains highly robust even under severe sparsity: at 5% density, performance remains strong and still surpasses operational estimators such as OPERA (CSI 0.323) by more than 24% in CSI. This demonstrates the model's ability to leverage limited point observations while still capturing localized rainfall structure. Figure 9 summarizes the performance trend, and Figure 10 illustrates a qualitative example showing the station inputs (5%, 30%, and 100%), radar observations, the resulting predictions by DropsToGrid, and SYNOP targets, along with the corresponding CSI values. As shown in the case study, with more input stations, the level of detail of DropsToGrid increases.

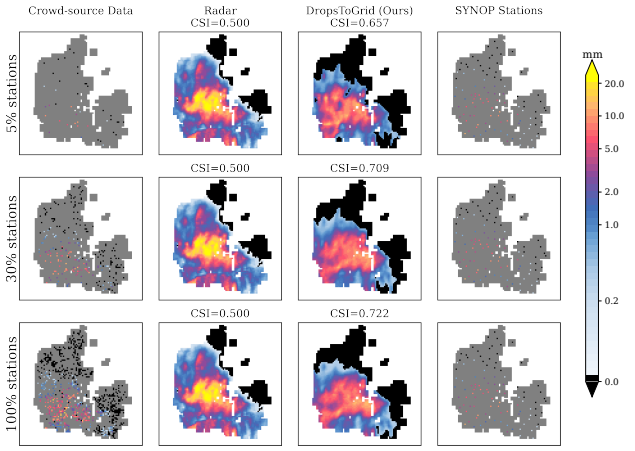

Figure 10. Qualitative example of station masked inputs, radar input, predictions, and targets.

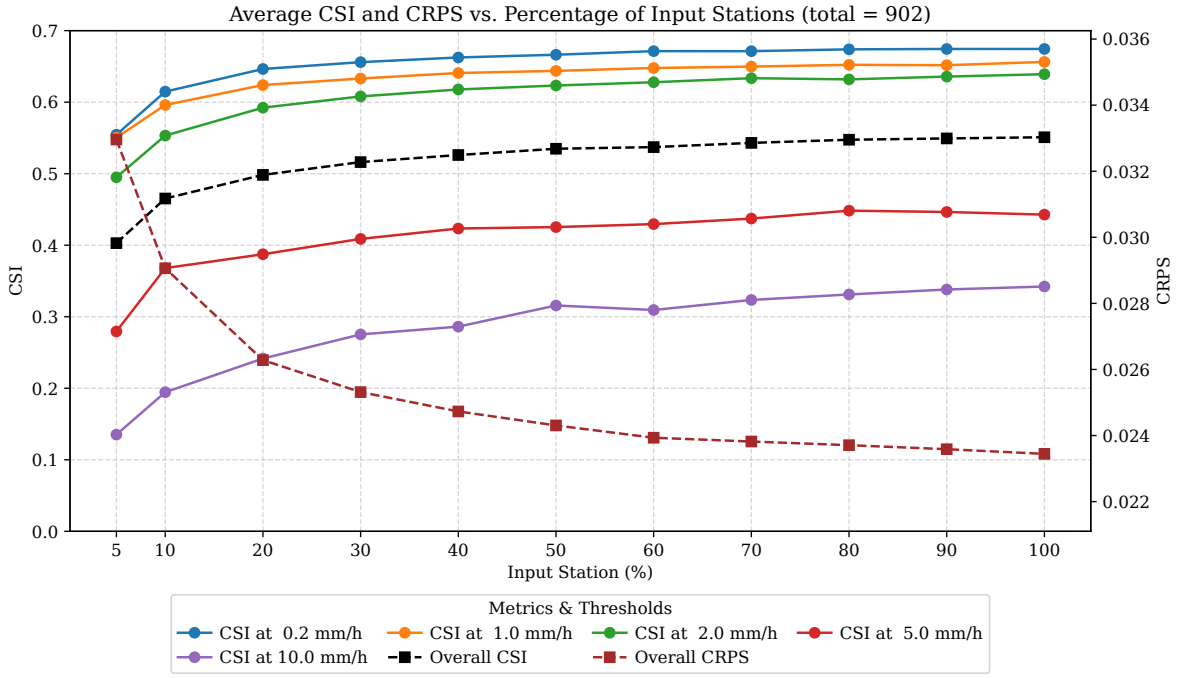

Figure 9. Average performance across station density levels.

14. Europe-wide densification

We evaluate DropsToGrid across the entirety of Europe for the full year of 2025, a period entirely unseen during training, encompassing a wide range of climatic and topographic conditions. This evaluation requires multiple sources: the WeatherUnderground PWS Network[4], OPERA radar, and GHCNh SYNOP stations[5], all providing at least EU-wide coverage. Table 14 presents results in which a model trained exclusively on Danish data (DropsToGrid_dk) is evaluated across Europe against both operational baselines and a learned baseline (ConvCNP_dk).

We further demonstrate scalability and improved performance when the same compact, parameter-efficient model is trained on all available EU data (DropsToGrid_eu). While training still uses patch-based inputs, inference in DropsToGrid is patch-free, achieved through matched TE-Transformer windows, which eliminates border artifacts. EU-wide inference is efficient, taking 0.20s for DropsToGrid versus 0.47s for ConvCNP, the latter being slower due to patch-based processing and 3D convolutions.

Climate baseline is excluded from EU-wide evaluation because it is only available for Denmark and utilizes SYNOP stations. Consequently, Table 14 reflects the limitations of global PWS and SYNOP networks, including lower data quality, sparser coverage, and heterogeneous temporal sampling compared to the Denmark-focused dataset (Table 1), leading to overall performance drops across baselines and learned models. Despite these challenges, DropsToGrid consistently outperforms all baselines. In contrast, ConvCNP trained in Denmark suffers from missing temporal attention and cross-attention mechanisms, limiting its ability to handle longer temporal gaps and heterogeneous spatiotemporal sources.

Overall, DropsToGrid exhibits robust generalization to unseen regions and seasons across Europe in 2025. It learns to suppress noise from PWS temporal observations, rely on radar in sparsely observed areas, and correct biases in radar-only regions by capturing systematic noise patterns. These capabilities allow DropsToGrid to substantially surpass operational baselines, regardless of network quality, coverage, or temporal heterogeneity.

[4] https://www.wunderground.com/pws/overview

[5] https://www.ncei.noaa.gov/products/global-historical-climatology-network-hourly

Table 14. Performance comparison across regions and station types for all of 2025. **Note:** PWS and SYNOP stations differ from those in the main paper, going from dense DK to sparser and noiser but wider EU coverage. With worse data quality and access, overall performance is lower, yet, DropsToGrid maintains similar improvements over baselines as in Table 1, demonstrating robust rainfall estimation.

Method	Denmark						Europe					
	PWS holdout			SYNOP Stations			PWS holdout			SYNOP Stations		
	CSI↑	FSS↑	MAE↓	CSI↑	FSS↑	MAE↓	CSI↑	FSS↑	MAE↓	CSI↑	FSS↑	MAE↓
OPERA	0.206	0.395	0.068	0.281	0.409	0.123	0.298	0.608	0.082	0.261	0.455	0.093
RainViewer	0.240	0.471	0.069	0.289	0.430	0.130	0.314	0.619	0.086	0.308	0.540	0.113
IMERG	0.182	0.392	0.121	0.232	0.351	0.216	0.199	0.473	0.120	0.205	0.426	0.148
ERA5	0.157	0.350	0.103	0.197	0.299	0.161	0.144	0.321	0.119	0.171	0.324	0.137
Climate	0.388	0.713	0.057	-	-	-	-	-	-	-	-	-
ConvCNP_dk	0.330	0.624	0.059	0.337	0.462	0.123	0.253	0.466	0.085	0.243	0.407	0.110
DropsToGrid_dk	**0.451**	**0.741**	**0.046**	*0.465*	*0.606*	**0.093**	*0.462*	*0.785*	*0.050*	*0.387*	*0.628*	**0.079**
DropsToGrid_eu	*0.444*	*0.729*	**0.046**	**0.472**	**0.614**	*0.096*	**0.476**	**0.802**	**0.049**	**0.400**	**0.646**	**0.079**

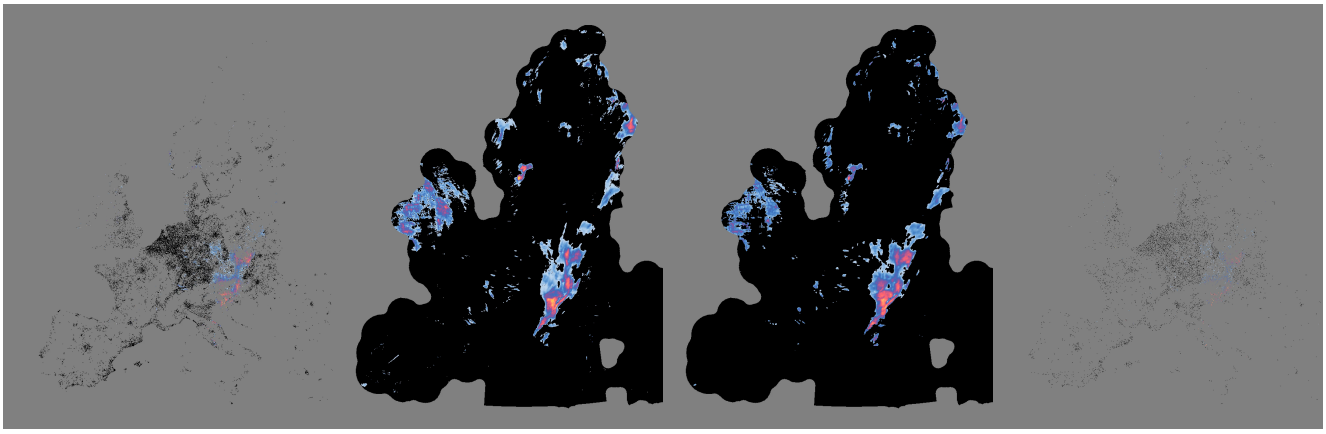

Figure 11. Sample EU densification. From left to right: input PWS stations, input radar, DropsToGrid prediction, and PWS holdout stations.

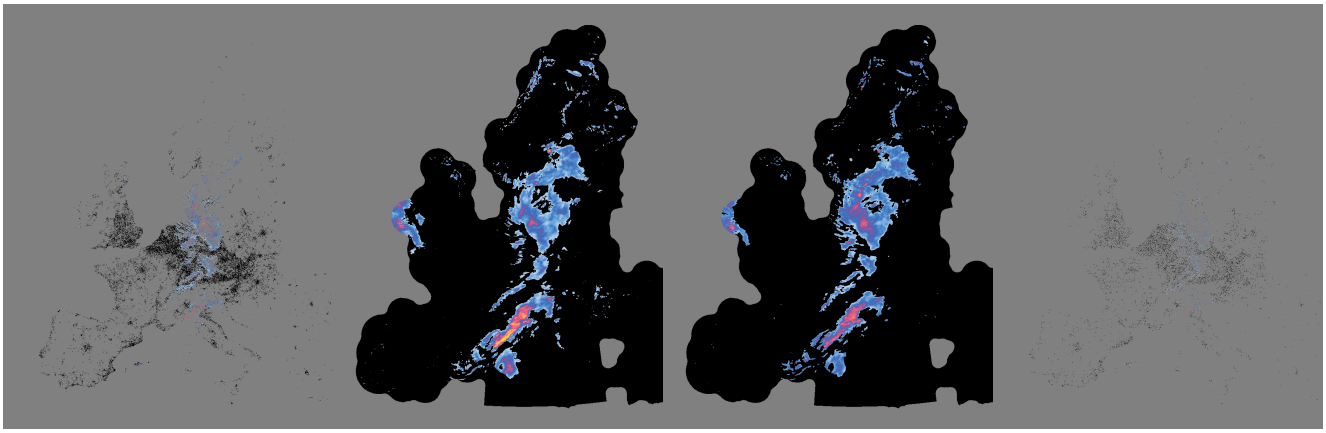

Figure 12. Sample EU densification. From left to right: input PWS stations, input radar, DropsToGrid prediction, and PWS holdout stations.