From research to applications
Examples of operational post-processing using machine learning

Maxime Taillardat
Météo-France / CNRM UMR 3589
maxime.taillardat@meteo.fr
ECMWF ML seminar series -
10/01/2020
1. Why ML-based ensemble PP?
2. Station-wise PP of surface Temperature
3. Gridded PP of hourly Rainfall
4. Conclusion and prospects
Motivation of post-processing ensemble forecasts

Météo-France’s 35-members global ensemble system (PEARP), 10 km resolution over France.

Observations and forecasts of 2-m temperature (T2m) at Lyon-Bron for the run of 1800UTC (different lead times)

- Run: 2012–10–26 Lead time: 12 h
  - Rank of the observed temperature: 18
  - Temperature (°C)

- Run: 2012–12–04 Lead time: 12 h
  - Rank of the observed temperature: 36
  - Temperature (°C)

- Run: 2013–05–12 Lead time: 36 h
  - Rank of the observed temperature: 36
  - Temperature (°C)
Motivation of post-processing ensemble forecasts

Let $Z = F(Y)$ (PIT statistic), $Z$ must verify

$$E(Z) = 1/2$$

$$\nabla(Z) = 12\text{var}(Z) = 1$$

- Need of a **simultaneous** correction of bias and dispersion
- Whatever the quality of the raw ensemble, post-processing improves forecast attributes (Hemri, 2014)
The state of the art

- Existing methods: Analogs, N Nets, Rank-based matching, CDF-matching, Member dressing, Bayesian Model Averaging, SAMOS...
- A review of techniques: Gneiting, 2014 and Vannitsem et al., 2020

The most (famous // widely-used) post-processing method:

- **Ensemble model output statistics (EMOS)** (Gneiting, 2005) also called Non-homogeneous Regression (NR)
  - fitting parameters linearly on predictors on some training period:

\[
f(y|x_1, \ldots, x_N) = \mathcal{N}(\mu = a_0 + \sum_{i=1}^{N} a_ix_i, \sigma^2 = b + cs^2)
\]

- \(y\) response variable, \(x_1, \ldots, x_N\) ensemble member values or any other predictor, \(s^2\) ensemble variance
\[ Y|X = w_1 \mathcal{N}(a_0 + b_0 \text{CTRL}, 0.1^2) + (1 - w_1) \mathcal{N}(a_1 + b_1 \text{ENS}, c_1 + d_1 \sigma_{\text{ENS}}^2) \]
Benefits of non-parametric post-processing

No assumptions on the weather variable you deal with

- Raw ensemble (not post-processed):

Non-parametric
Benefits of non-parametric post-processing

No assumptions on the weather variable you deal with

- Raw ensemble (not post-processed):

- EMOS post-processing:

Non-parametric
Benefits of non-parametric post-processing

No assumptions on the weather variable you deal with

- Raw ensemble (not post-processed):

- EMOS post-processing:

Non-parametric

- Post-processing possible results:
Benefits of non-parametric post-processing

No assumptions on the weather variable you deal with

- Raw ensemble (not post-processed):

  ![Hyena](image1)

- EMOS post-processing:

  ![Camel](image2)

Non-parametric

- Post-processing possible results:

  ![Hyena](image3)
  ![Camel](image4)
  ![Camel](image5)
  ![Camel](image6)
  ![Giraffe](image7)
Method of PP employed: QRF, another way to find analogues

$$\hat{P}(Y \leq y | X = x) = \frac{1}{N} \sum_{i=1}^{N} w_i(x)1\{Y_i \leq y\}$$

**Pros**

- No assumptions on the target variable
- Self-selection of the most useful predictors, interpretable
- Hyperparameters tuning quite easy and stable over locations vs. other ML techniques
Best-of post-processing possible outputs

12-h forecast at MELUN the 2013–03–12

6-h forecast at CARCASSONNE the 2014–02–10

27-h forecast at BOULOGNE-SUR-MER the 2014–09–01

6-h forecast at PARIS–LE–BOURGET the 2014–03–11
Outline

1. Why ML-based ensemble PP?
2. Station-wise PP of surface Temperature
3. Gridded PP of hourly Rainfall
4. Conclusion and prospects
Observations available on 2000 stations locations across Western Europe

Raw model resolution: 10km (not native grid)

Operational challenges

- available "homogeneous" PEARP archive: 2 years
- 1 forest/station/lead time/init. time: 260000 models, 1.2TB...

Goal

- Station-wise post-processing with ECC
- Target resolution: 1km (Downscaling step), 4.000.000 grid points

- QRF calibration: 12’ on 3 HPC nodes (1 node: 128 cores / 256GB mem.)
- ECC+regression-kriging (+ edition of 1400 GRIB files): 12’ on 3 HPC nodes

NB: 35-member PEARP forecast run: 60’ on 205 nodes
Towards high resolution temperature fields: Illustration
Results of QRF station-wise PP

T2m CRPS on EURW1S100 Stations

T2m PIT mean on EURW1S100 Stations

T2m PIT variance on EURW1S100 Stations

Maxime Taillardat
11/25
Outline

1. Why ML-based ensemble PP?
2. Station-wise PP of surface Temperature
3. Gridded PP of hourly Rainfall
4. Conclusion and prospects
How to verify/evaluate ensemble forecasts?

- Differ from deterministic forecasts
- A "point" (eg. for one day) verification is a nonsense
- It has to be **statistical**

**Several attributes sought**

- **Reliability**
  - Accordance between forecasted probabilities and observed frequencies of an event and/or exchangeability between observations and ensemble members
- **Resolution/Discrimination** (Bröcker, 2015 e.g.)
  - Ability to differ from a climatological forecast
- **Sharpness**
  - Getting the least dispersed forecast

"Maximizing sharpness subject to calibration" (Gneiting et al. 2006)
How to verify/evaluate ensemble forecasts?

"Maximizing value for extremes subject to a good overall performance"
"Here comes the rain again..."

Hourly rainfall

- A lot of zeros observed and (well) forecast.
- Point mass, asymmetry
- Extremes

A semi-parametric approach

Use QRF outputs to fit a distribution which would:

- Model jointly low, moderate and heavy rainfall
- Be flexible
- Use of an Extended GP distribution (EGP3) (Papastathopoulos and Tawn, 2013; Naveau et al., 2016; Tencaliec et al. 2019)
A semi-parametric approach

Our final distribution is:

\[ G(x) = f_0 + (1 - f_0) \left[ 1 - \left( 1 + \frac{\xi x}{\sigma} \right)^{-\frac{1}{\xi}} \right]^\kappa \]

Strategy

1. Run QRF to get \( \hat{F}(y|X = x) = \hat{P}(Y \leq y|X = x) \)
2. Keep the probability of no rain \( \hat{f}_0 = \hat{P}(Y = 0|X = x) \) from QRF outputs
3. Estimate \( (\hat{\kappa}, \hat{\sigma}, \hat{\xi}) \) from non-zero QRF quantiles

What we expect:
A semi-parametric approach

Our final distribution is:

\[ G(x) = f_0 + (1 - f_0) \left[ 1 - \left( 1 + \frac{\xi x}{\sigma} \right)^{-\frac{1}{\xi}} \right]^\kappa \]

Strategy

1. Run QRF to get \( \hat{F}(y|X = x) = \hat{P}(Y \leq y|X = x) \)
2. Keep the probability of no rain \( \hat{f}_0 = \hat{P}(Y = 0|X = x) \) from QRF outputs
3. Estimate \( (\hat{\kappa}, \hat{\sigma}, \hat{\xi}) \) from non-zero QRF quantiles

What we expect:

- QRF possible output
A semi-parametric approach

Our final distribution is:

\[ G(x) = f_0 + (1 - f_0) \left[ 1 - \left( 1 + \frac{\xi x}{\sigma} \right)^{-\frac{1}{\xi}} \right]^\kappa \]

Strategy

1. Run QRF to get \( \hat{F}(y|X = x) = \hat{P}(Y \leq y|X = x) \)
2. Keep the probability of no rain \( \hat{f}_0 = \hat{P}(Y = 0|X = x) \) from QRF outputs
3. Estimate \( (\hat{\kappa}, \hat{\sigma}, \hat{\xi}) \) from non-zero QRF quantiles

What we expect:

- QRF possible output
- After "EGP TAIL"
Operational framework for hourly rainfall

- PEAROME (16 members, 2.5 km), lead times from 1 to 45 hours, 4 times/day
- 300,000 grid points over France
- Observations: Radar+rain gauges ANTILOPEJP1H (1 km)
- Archive: 2 years, 20TB
- Restore scenarios (post-processed members).

**Challenges**

- 300,000*45*4 models? Too many I/Os
- Variable selection: from 50 to 20
- High resolution: Double penalty issue
**Data pooling:** We consider high res. errors homogeneous on 10km boxes (spatial penalty). PP is made on these HCA: number of statistical models reduced by a factor 25. (14000 HCA)

**Data boosting:** As observation is at 1km, observation is a distribution. The length of the training sample is inflated by a factor 5.

For 1 init. time: 600.000 models, 600GB.
What sort of members do we want?

- Schaake Shuffle (SS) and MD-SS (see e.g. Clarke, 2005; Scheuerer, 2018)
  - We need an observations archive, we lose the model "signature"

- Ensemble Copula Coupling-like methods (ECC) (see e.g. Schefzik et al., 2013; Ben Bouallègue et al., 2017)
  - Using (potentially wrong) physical structures of the raw ensemble
ECC and rainfall: not so simple...

Example of bc-ECC ($b = \infty, c = 1$) for 2-member (M) ensemble in a 3 grid Point linear HCA.

<table>
<thead>
<tr>
<th>In the HCA</th>
<th>gP1M1</th>
<th>gP2M1</th>
<th>gP3M1</th>
<th>gP1M2</th>
<th>gP2M2</th>
<th>gP3M2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw values</td>
<td>2</td>
<td>2</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>HCA Calibrated values</td>
<td>0</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>$b$-ECC and average</td>
<td>5.5</td>
<td>5.5</td>
<td>7</td>
<td>2</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>Is rain in M in c gP around?</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>no</td>
<td>yes</td>
<td>–</td>
</tr>
<tr>
<td>Final values</td>
<td>5.5</td>
<td>5.5</td>
<td>7</td>
<td>0</td>
<td>2</td>
<td>5</td>
</tr>
</tbody>
</table>

Example of bc-ECC ($b = \infty, c = 1$) for 2-member (M) ensemble in a 3 grid Point linear HCA.

Bootstrapped-Constrained Ensemble Copula Coupling (bc-ECC)

We do ECC many times (here 250 times by HCA) and average values:

- If raw zeros > calib. zeros: smallest non-zero calib. rainfall are assigned and averaged on raw zeros
- a raw zero becomes a non-zero member IF there is a raw non-zero member in a 2 grid point neighborhood
ECC + post-processing visualization

2 PP members (left) with their associated raw members (right)
Rain discrimination results

Looking at (lead time-averaged) CRPS...

- Hourly rainfall: from **0.118** (raw) to **0.0790** (PP)
- Daily rainfall (After bc-ECC): from **2.17** (raw) to **2.11** (PP)

- QRF EGP TAIL calibration: 12’ on 6 nodes
- bc-ECC+ 1920 GRIB Edition: 3’ on 6 nodes

NB: 16-member PEAROME forecast run : 64’ on 129 nodes
Outline

1. Why ML-based ensemble PP?
2. Station-wise PP of surface Temperature
3. Gridded PP of hourly Rainfall
4. Conclusion and prospects
A matter of triptychs

For a good R2O process

Dialog should be straightened (enabled ?) among communities.
Ready for the cultural change? (cf. A. Arribas talk)

PP: escape from the NWP model land

DL for local PP... And for direct 2D PP fields

- Promising architectures/tools: CNN/U-Net/RNN, embeddings in space-lead time-season-topography...
- (Non-exhaustive) list of examples:
  - **Temperature**: Rasp and Lerch, 2018
  - **Wind speed**: Bremnes, 2020; Bhend et al., 2020; Veldkamp et al., 2020; Candido et al., 2020
  - **Cloud cover**: Dupuy et al., 2020; Bhend et al., 2020
  - **T850, Z500**: Grönquist et al., 2020

Verification issues

- Quality vs. value
- Local (log score) vs. Distance-sensitive (CRPS) scoring rules
- Go beyond averaging scoring rules? (see e.g. Taillardat et al., 2019)
  Example: ECMWF headline score of %CRPS $> 5K$
Some references

- PP and R2O

- Quantile Regression Forests and PP

- DL-based PP

- Ensembles’ Verification issues