CBPR Parameter Estimation Program User's Guide

by

Sunghee Kim, PhD

NOAA/NWS Office of Water Prediction (OWP) sunghee.kim@noaa.gov
Lynker, Leesburg, VA 20176 skim@lynker.com

Feb 29, 2024

Table of Contents

1 Overview

This document describes how to run the R script for conditional bias-penalized regression (CBPR), *cbpr.R*.

It is assumed that the reader is familiar with the Meteorological Ensemble Forecast Processor Parameter Estimation Program (MEFPPE, NWS 2019), the Meteorological Ensemble Forecast Processor (MEFP, Schaake et al. 2007, Wu et al. 2011, NWS 2017), the Global Ensemble Forecast System (GEFS) and its reforecast products (Guan et al. 2022, Hamill et al. 2022), Kim et al. (2023) and Kim (2024). CBPR is for precipitation only and does not deal with temperature. All precipitation hindcasts are forced by the GEFS ensemble mean precipitation reforecast.

Fig 1 shows the relationship among *cbpr.R*, MEFPPE and MEFP. In the figure, *cbpr.R*, denoted as CBPR, inputs the forecast-observation pairs for all canonical events (CE) in *<loc>.precipitation.mefp.parameters.tgz*, where *<loc>* is the location name, and updates the above *.tgz* file with the α-appended MEFP parameters (see Kim 2024 for explanation of α). The rest of the hind- or forecast process is the same as when CBPR is not involved. In the figure, *preproc.R*, denoted as Preproc, rewrites the forecast-observation pairs provided in a binary file to a *.csv* for ingestion by *cbpr.R*.

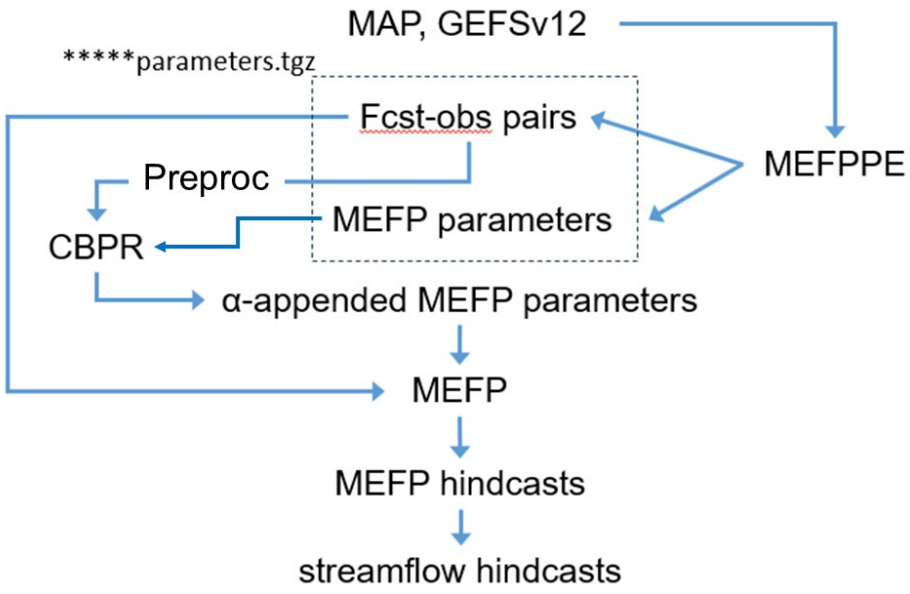

Fig 1. Relationship among CBPR, MEFPPE and MEFP.

The R script, *cbpr.R*, performs optimization of λ (Kim et al. 2023, Kim 2024) valid for 73 days (once every 5 days) of the 365-day year and updates *<loc>.precipitation.mefp.parameters.tgz*. The resulting *.tgz* file may be used for hindcasting or operationally at the RFCs. All CBPR-related files are located on CEE under */gefsv12/Longterm_storage/OWPproj/cbpr/*. The R script, *cbpr.R*, is located under */cbpr/src*. Fig 2 shows the directory structure.

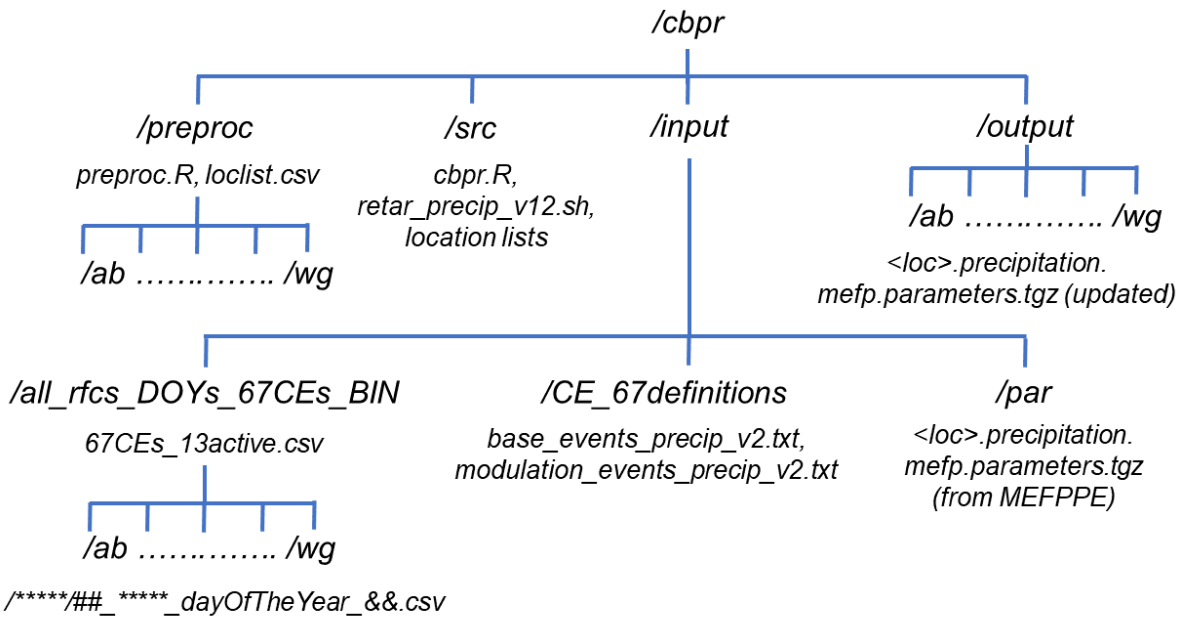

Fig 2. Directory structure associated with the CBPR script. The data generated during the final testing of the scripts for this user's guide are kept under */cbpr/input/all_rfcs_DOYs_67CEs_BIN/* and */cbpr/input/par* as examples for the user.

2 Adaptable parameters

CBPR employs the following adaptable parameters to optimize performance and to maximize consistency with MEFPPE (see also Kim et al. 2023, Kim 2024).

bias20_tol
crpss21_tol

These two parameters prescribe the maximum (unconditional) multiplicative bias (≥ 1) and the maximum percent degradation (≤ 0) of (unconditional) CRPSS tolerated in CE-specific optimization of λ (see Kim et al. 2023 for details). The same settings are used for all CBPR-active CEs (see Kim 2024 for explanation). The default values are 1.20 for *bias20_tol* (i.e., a wet bias of up to 20% tolerated) and -0.03 for *crpss21_tol* (i.e., degradation of CRPSS of up to 3% tolerated).

In the variable names above, the numbers *0*, *1* and *2* are associated with the observation, the ordinary least squares regression (OLSR) prediction and the CBPR prediction, respectively. Recall in Kim et al. (2023) that OLSR is used in the RFC-operational MEFP. Hence, *bias20* represents the multiplicative bias in the (unconditional) mean of the CBPR prediction in reference to the observation and *crpss21* represents the (unconditional) CRPSS of the CBPR prediction in reference to the OLSR prediction.

The above two parameters control the Type-I vs. Type-II error tradeoff and hence are very important as explained in detail in Kim et al. (2023). In one extreme of *bias20_tol=1* and *crpss21_tol=0*, one is not at all willing to increase Type-I error for reduced Type-II error. Hence, CBPR effectively becomes OLSR. In the other extreme of *bias20_tol*$\rightarrow \infty$ and *crpss21_tol*$\rightarrow -\infty$, one is willing to increase Type-I error freely to reduce Type-II error. Hence, CBPR minimize Type-II error with no regard to Type-I error when optimizing λ.

numcut

This parameter prescribes the maximum number of incremental increases in the cutoffs for cutoff-dependent optimization of λ (see Kim et al. 2023 for details). The default is 39. The choice of the increment by which the cutoff is increased in each iteration depends on the precipitation duration as shown in the table below.

Table 1. Duration-dependent increments used in the optimization of λ.

Duration (hrs)	Increment (IN)
6	0.05
12	0.1
24	0.2
48	0.2
72	0.3
> 95	0.4

thresh

This parameter specifies the minimum detectable precipitation in inches. The default is 0.01 (IN) (per Hank Herr). This parameter is invoked only when the MEFP parameters are estimated via the function *params_estim( )* in *cbpr.R* independently of MEFPPE.

shift

This parameter prescribes the shift to be applied to both observed precipitation and the GEFS ensemble hindcasts (per Hank Herr). The default is 0.249/25.4 (IN). See Kim (2024) for details.

The following 3 adaptable parameters prescribe the minimum sample size. The following naming convention is used:
10: forecast precipitation is positive (i.e., above the threshold) and observed precipitation is zero (i.e., below the threshold),
01: forecast precipitation is zero and observed precipitation is positive, and
11: both forecast and observed precipitation amounts are positive.

ss10_min

This parameter prescribes the minimum sample size for gamma/Weibull 10, i.e., for modeling the probability distribution of positive forecast precipitation when the verifying observation is zero using the 2-parameter gamma/Weibull distribution. The default is 5.

ss01_min

This parameter prescribes the minimum sample size for gamma/Weibull 01, i.e., for modeling the probability distribution of positive observed precipitation when the forecast precipitation is zero using the 2-parameter gamma/Weibull distribution. The default is 5.

In MEFPPE-MEFP, gamma/Weibull 01 is not modeled (per Hank Herr) and hence the user may ignore this adaptable parameter.

ss11_min

This parameter prescribes the minimum sample size for gamma/Weibull 11, i.e., for modeling the marginal probability distributions of positive forecast precipitation and positive observed precipitation using the 2-parameter gamma/Weibull distribution. The default is 30.

Note that the distribution modeling in MEFPPE involving Weibull has inconsistencies with the CBPR script as described in detail in Kim (2024). For the implications and potential impact to precipitation hindcasts, the reader is referred to Kim (2024).

ss_cbpr_min

This parameter prescribes the minimum sample size for the cutoff-dependent optimization of λ. The default is 7. A smaller *ss_cbpr_min* allows the cutoff to go higher and hence reflects the conditional bias for the largest verifying observations but at the expense of increased sampling uncertainty. A larger *ss_cbpr_min*, on the hand, suppresses sampling uncertainty but at the expense of limiting how high the cutoff can go.

rel_tol

This parameter prescribes the relative accuracy requested for numerical integration using the R function, *integrate( ).* The default is 0.01. The integration method used is adaptive quadrature.

igamma_internal

This parameter prescribes how the MEFP parameters are estimated for the optimization of λ. If they are to be estimated within *cbpr.R* independently of MEFPPE, set *igamma_internal* to 1. If they are to be retrieved from MEFPPE, set *igamma_internal* to 0. The default is 1. For details on the discrepancies in the MEFP parameters estimated by MEFPPE vs. those estimated by the R script, their potential sources, implications and potential impact, see Kim (2024).

iopt

This parameter prescribes the distribution model and the parameter estimation method to be used. This parameter is valid only if *igamma_internal* is set to 1.

There are currently 4 options available for fitting the marginal distribution of positive forecast precipitation when the verifying observed precipitation is positive ($D_X(x)$ in Wu et al. 2011), the marginal distribution of positive observed precipitation when the forecast precipitation is positive ($D_Y(y)$) and the marginal distribution of positive observed precipitation when the forecast precipitation is positive (($G_Y(y)$ in Wu et al. 2011) (see Table 2). The default is 1 for SERFC and 3 for all other RFCs (see Kim 2024 for details). The use of $iopt = 0$ is currently not recommended.

Table 2. Options for modeling and parameter estimation for $D_X(x)$, $D_Y(y)$ and $G_Y(y)$.

iopt	Distr. model	Para. estim. method	R pckg. used	Reference
0	2-para. Weibull	L-moment	lmomco	Asquith (2023)
1	2-para. Weibull	Max. likelihood	MASS	Venerables and Ripley (2002)
2	2-para. gamma	Max. likelihood	MASS	Venerables and Ripley (2002)
3	2-para. gamma	L-moment	lmomco	Asquith (2023)

jopt

For fitting the marginal distribution of positive forecast precipitation when the verifying observed precipitation is zero (($G_X(x)$ in Wu et al. 2011), MEFPPE uses gamma only and hence there are only 2 options available (see Table 3). As explained in detail in Kim (2014), MEFPPE uses 3-parameter gamma (or Pearson Type III) but with the shift parameter replaced by a fixed value, which is equivalent to 2-parameter gamma. The default for $jopt$ is 3 for all RFCs.

Table 3. Options for modeling and parameter estimation for $G_X(x)$.

jopt	Distr. model	Para. estim. method	R pckg. used	Reference
2	2-para. gamma	Max. likelihood	MASS	Venerables and Ripley (2002)
3	2-para. gamma	L-moment	lmomco	Asquith (2023)

3 Running *cbpr.R*

This section provides step-by-step instructions on running *cbpr.R*.

1) Configure the 67-CE definition for MEFPPE

   The CBPR-aided MEFP does not use the existing CE definitions but uses the newly-developed definition consisting of 67 CEs of which 13 are CBPR-active (see Kim 2024 for details). Note that 67 refers to the number of CEs within the first 14 days of lead time. To run the CBPR-aided MEFP, it is necessary first to configure the new CE definition in MEFPPE and newly estimate the MEFP parameters.

To configure the 67-CE definition in MEFPPE:

- Navigate to */cbpr/input/ CE_67definitions/,*
- Copy *base_events_precip_v2.txt* and *modulation_events_precip_v2.txt* (these names, including the version number, are hard-coded in MEFPPE and hence cannot be changed) to *<mefppe_sa>/Models/hefs/mefppeRunArea/import/control/*, where *<mefppe_sa>* is RFC-dependent, and
- Launch MEFPPE.

If the new definition is correctly configured, the user should see 50 base and 42 modulation events in the MEFPPE GUI where 92 (=50+42) represents the total number of CEs over the entire forecast horizon of 365 days.

2) Run MEFPPE to estimate the MEFP parameters. Copy the resulting *<loc>.precipitation.mefp.parameters.tgz* to */cbpr/input/par*.

3) Prepare/Verify input files

   - Verify that the location lists, /cbpr/src/*list_ab1* through /cbpr/src/*list_wg1*, exist.
   - Verify that the CE definition table (see Kim 2024 for details), /cbpr/*input/all_rfcs_DOYs_67CEs_BIN/67CEs_13active.csv,* exists.
   - Verify that the fcst-obs pairs exist under /cbpr/input*/all_rfcs_DOYs_67CEs_BIN* as shown in Fig 2. For example, the fcst-obs pairs for BSRC1HOF in CNRFC within the 61-day sampling window centered at DOY of 361 are in */cbpr/input/all_rfcs_DOYs_67CEs_BIN/cn/BSRC1HOF/cn_BSRC1HOF_doy_361.csv.*

     The .csv files under */cbpr/input/all_rfcs_DOYs_67CEs_BIN* were generated from *GEFSv12SourceModelParameters.precip.events.bin* in *<loc>.precipitation.mefp.parameters.tgz* via *preproc.R*.

     If the user wants to newly generate the fcst-obs pairs from the *.bin* file, do the following:

     - Run MEFPPE to generate *<loc>.precipitation.mefp.parameters.tgz* for all locations that the user wants to generate the fcst-obs pairs for.
     - Copy the resulting *<loc>.precipitation.mefp.parameters.tgz* to */cbpr/input/par*.
     - Navigate to */cbpr/preproc/*.
     - List the desired locations in *loclist.csv*. Each row should contain the RFC id *<rfc>* and the location id *<loc>*, separated by a comma.
     - Run *preproc.R*,
     - Move the existing *.csv* files under */cbpr/input/all_rfcs_DOYs_67CEs_BIN/<rfc>* to some other location of the user's choice so that the existing fcst-obs pairs are not overwritten.
     - Copy the output files under */cbpr/preproc/out/<rfc>* to */cbpr/input/all_rfcs_DOYs_67CEs_BIN/<rfc>*.

4) If the user wants to change the CBPR-active CEs, modify
/cbpr/all_rfcs_DOYs_67CEs_BIN/67CEs_13active.csv.
The last column in the above file indicates whether the CE is CBPR-active (1) or not (0).

5) If the user wants to change any of the CBPR adaptable parameters (see Section 2), edit *cbpr.R* and make changes.

6) Edit *cbpr.R*, verify the paths for input and output and change if necessary.

7) Run *cbpr.R.*

To generate the updated *<loc>.precipitation.mefp.parameters.tgz* files for all 149 locations in 13 RFCs in a batch mode, use *run_all* which runs the following 24 batch jobs:

```
R CMD BATCH -ab1 cbpr.R Rout_ab1 &
R CMD BATCH -ap1 cbpr.R Rout_ap1 &
R CMD BATCH -cb1 cbpr.R Rout_cb1 &
R CMD BATCH -cb2 cbpr.R Rout_cb2 &
R CMD BATCH -cn1 cbpr.R Rout_cn1 &
R CMD BATCH -cn2 cbpr.R Rout_cn2 &
R CMD BATCH -cn3 cbpr.R Rout_cn3 &
R CMD BATCH -cn4 cbpr.R Rout_cn4 &
R CMD BATCH -cn5 cbpr.R Rout_cn5 &
R CMD BATCH -lm1 cbpr.R Rout_lm1 &
R CMD BATCH -lm2 cbpr.R Rout_lm2 &
R CMD BATCH -ma1 cbpr.R Rout_ma1 &
R CMD BATCH -mb1 cbpr.R Rout_mb1 &
R CMD BATCH -nc1 cbpr.R Rout_nc1 &
R CMD BATCH -ne1 cbpr.R Rout_ne1 &
R CMD BATCH -ne2 cbpr.R Rout_ne2 &
R CMD BATCH -nw1 cbpr.R Rout_nw1 &
R CMD BATCH -nw2 cbpr.R Rout_nw2 &
R CMD BATCH -nw3 cbpr.R Rout_nw3 &
R CMD BATCH -nw4 cbpr.R Rout_nw4 &
R CMD BATCH -nw5 cbpr.R Rout_nw5 &
R CMD BATCH -oh1 cbpr.R Rout_oh1 &
R CMD BATCH -se1 cbpr.R Rout_se1 &
R CMD BATCH -wg1 cbpr.R Rout_wg1
```

The above runs may take up to about 6 hours per location, depending on the load on the computing system. The elapsed time may be found in the last line of *Rout_%%%*
where %%% is the 3-character location list identifier, ab1 through wg1.

8) Once all runs are successfully completed, verify that the updated MEFP parameter files, */cbpr/output/##/<loc>.precipitation.mefp.parameters.tgz*, have been produced, where ## is

the 2-character RFC name in lowercase. These *.tgz* files may be used for hindcasting or for operational use at the RFCs.

References

Asquith W.H., 2023. lmomco—L-moments, censored L-moments, trimmed L-moments, L-comoments, and many distributions. R package version 2.4.11.

Guan, H., Y. Zhu, E. Sinsky, B. Fu, W. Li, X. Zhou, X. Xue, D. Hou, J. Peng, M. M. Nageswararao, V. Tallapragada, T. M. Hamill, J. S. Whitaker, G. Bates, P. Pegion, S. Frederick, M. Rosencrans, R. Kumar, 2022. GEFSv12 reforecast dataset for supporting subseasonal and hydrometeorological applications. Monthly Weather Review. 10.1175/MWR-D-21-0245.1.

Hamill, Thomas M. et al., 2022. The Reanalysis for the Global Ensemble Forecast System, Version 12. 150(1). https://doi.org/10.1175/MWR-D-21-0023.1

Kim, S., A. Jozaghi, and D.-J. Seo, 2023. Improving ensemble forecast quality for heavy-to-extreme precipitation for the Meteorological Ensemble Forecast Processor via conditional bias-penalized regression, manuscript to be submitted to J. Hydrol.

Kim, S., 2024. Comparative Evaluation of CBPR-Aided MEFP: Summary, Findings and Technical Recommendations, Technical Document, Lynker, Leesburg, VA, 87pp.

National Weather Service, 2016. MEFPPE Configuration Guide, NOAA/NWS/Office of Water Prediction, Silver Spring, MD. Accessed 27 June, 2019. [Available online at https://vlab.ncep.noaa.gov/documents/207461/1893010/MEFPPEConfigurationGuide.pdf]

National Weather Service, 2017. Meteorological Ensemble Forecast Processor (MEFP) User's Manual, NOAA/NWS/Office of Water Prediction, Silver Spring, MD. Accessed 27 June, 2019. [Available online at https://vlab.ncep.noaa.gov/documents/207461/1893026/MEFP_Users_Manual.pdf]

Schaake J., J. Demargne, M. Mullusky, E. Welles, L. Wu, H. Herr, X. Fan, and D.-J. Seo, 2007. Precipitation and temperature ensemble forecasts from single-value forecasts, 4, Hydrology and Earth System Sciences, 655-717.

Venables W.N., B.D. Ripley, 2002. Modern Applied Statistics with S, Fourth edition. Springer, New York. ISBN 0-387-95457-0, https://www.stats.ox.ac.uk/pub/MASS4/.

Wu, L., D.-J. Seo, J. Demargne, J. Brown, S. Cong and J. Schaake, 2011. Generation of ensemble precipitation forecast from single-valued quantitative precipitation forecast via meta-Gaussian distribution-based models, J. Hydrol., 399(3-4), 281-298.