We thankfully acknowledge the Handling Editor and Reviewers for their very helpful comments on the revised manuscript, which enabled us to further polish the presentation of our study. We addressed all comments by Reviewer, as illustrated in our reply here below.

> *We report each Reviewer's comment using this format;*

While our reply is formatted in this way.

All modifications are highlighted in a copy of the manuscript with tracked changes attached to our reply (end of this document). Page and lines number also refer to the manuscript with highlighted changes.

> *Reviewer#1*
>
> *All the issues reported in the first review step have been addressed by the authors in the new version of the paper, now titled "Geostatistical prediction of flow-duration-curves in an index-flow framework". I believe the work has been improved, so I recommend it for publication in HESS.*
>
> *I have only one minor technical note. It is not clear what "different combinations of drainage areas" means (line 12 on page 15 and fig.4) when the authors describe the binning procedure. I would change the sentence to better specify how the area is used to create bins.*

The Reviewer is right; the description of the binning procedure and the illustration of Figure 4 can be improved. Therefore we modified the text and we clarified the meaning of each element of Figure 4 in the figure caption.

Original text:
[…] Figure 4 shows the differences between fitted variograms by either using no bins (i.e. point variogram) or by binning groups of pairs of catchments with different combinations of drainage areas. […]

P. 15, lines 19-25, modified text:
Figure 4 illustrates the comparison between some selected bins of the sample variogram and the regularized semivariance for those bins (see also Skøien et al., 2014 and Figure 4 therein). The numbers in the legend refer to different combinations of catchment areas, so that, e.g., 300 vs. 75 means the regularized semivariogram as a function of distance for two catchments of size ~300 and ~75 km^2, respectively; while the solid lines represent a regularized semivariogram of equally sized catchments (~300 km^2).

Original caption of Figure 4:
Empirical and theoretical semivariograms from TND_1 values. Black line shows the fitted point variogram. Colour markers and lines show empirical and fitted variograms (empirical variograms are computed by binning catchment pairs for different combinations of catchment areas, e.g. ~300 km^2 vs ~75 km^2).

Modified caption of Figure 4:
Sample variograms (points) and regularized semivariances as function of distance and area. The solid blue line represents regularized semivariance of equally sized catchments (~300 km^2), dotted lines combinations of catchments sizes (see also Figure 4 in Skøien et al., 2014).

We also included among the references:
Skøien, J. O., G. Blöschl, G. Laaha, E. Pebesma, J. Parajka, A. Viglione, rtop: An R package for interpolation of data with a variable spatial support, with an example from river networks, Computers & Geosciences, 67, 180-190, ISSN 0098-3004, 2014.

Reviewer#2

General comments

This is the second time I review the paper and it is nice to see that it has been improved from the last version. However, with regards to the presentation I would still urge the authors to revise the text thoroughly to improve the clarity in the presentation of their work. I have given only a few technical comments below in this respect.

With regards to the content, the main thing that needs to be addressed is a better explanation of the fitting of the variogram and Figure 4, see comments below.

We thankfully acknowledge Reviewer#2 for her/his 2nd assessment. Reviewer #1 requested a clarification of the meaning of Figure 4 also.

Specific comments

P2, line 25-26. But the low flow performance was not so good for the un-nested case, wherefore this conclusion should be modified.

The prediction of the low-flow regime in the un-nested case for the proposed procedure (Leave Nested Out Cross-Validation, LNOCV) is still as good as, or better than, the other considered regional model (that do not neglect the information coming from nested catchment). However, the revised discussion of the un-nested case explicitly points out the slight decrease of cross-validation performance that can be observed for low-flow regime (see P. 24, line 25-28).

Also exclude emphasis words like "very".

We revised the manuscript according to this suggestion.

I would also suggest focusing the abstract more on what was done in the study. I would suggest presenting the three study aims and the corresponding results, including more information about the TND indices and the sensitivity analyses that were done and reducing the general sentences about FDCs in the beginning.

We entirely restructured the abstract (see the abstract in the tracked-changes version of the manuscript).

P14. This section should refer to the new information in the supplementary material.

According to this suggestion we included a reference to supplementary material (see P. 15, line 7-8).

P15. L8-9. It should be specified how many bins were used. From Figure 4 it looks like there are fewer bins than the 5 parameters in the variogram fitted to them, which seems like overfitting??

Overfitting is not the case here, as each bin includes several empirical values of semivariance and the five parameters of the one and only theoretical point variogram used by the procedure are fitted in a single optimization procedure by looking simultaneously at all the regularised empirical sample variograms computed for various classes of catchment sizes (coloured points in Figure 4, see also detailed description in Skøien et al., 2014). However, we did not develop the fitting procedure -and its detailed illustration is clearly out of the scope of our manuscript-, which was originally presented in Skøien et al. (2006) and is now implemented in the R-package rtop and clearly illustrated in the associated vignette (Skøien et al., 2014) that is now cited in the text.

> *You should also comment on how well the variograms were fitted to the binned data, the fit does not seem to be that good from Figure 4?*

The fitting is analogous to other Top-kriging applications see e.g. Castiglioni et al., (2011) and Figure 3 therein; Skøien et al. (2014) and Figure 4 therein. We do not believe that commenting on this is beneficial to our paper (i.e. the study focuses on the prediction of FDCs, this is what we are really interested in).

> *Additionally for Figure 4:*
>
> *• It should be "semivariance" on the y-axis and the x and y axes should go down to zero so the whole variogram can be seen. What is the difference between the red diamonds (and the other markers) of varying size? The dark red diamond is very difficult to discern from the bright red.*

The Reviewer is right, the *y*-axis label is not accurate, it should say "semivariance". We increased the axes range as requested, we also clarified the meaning of the plotted lines and points. Size of the markers is proportional to the number of pairs within a bin (i.e. the higher number in the bin the bigger marker). Unfortunately marker types and colours are not modifiable from the package settings.

> *• The semivariance cloud could be shown in the background so the appropriateness of the binning used can be assessed and also the quality of the fit.*

We believe that there is already a lot of information in the plot, adding the empirical semivariance associated with the point variogram would reduce significantly the readability of the plot.

> *• Also it is unclear what 300km2 vs 75 km2 means, is it referring to catchment pairs with areas in the intervals 1-300km2 vs 1-75 km2?*

Please refer to our reply to Reviewer#1 and changes to the text and figure caption.

> *P15 L22, could specify again that d is duration*

Done

> *P19, line 19-25. "A comprehensive error index…" but this seems to be just the summed absolute errors at the selected points? How were these points chosen? In the same way as on P15? The authors could consider moving the whole paragraph about the evaluation methods on page 15 and 16 to this section about the performance indices so that the prediction method*

is discussed first and the evaluation of it later to keep a clearer structure of the methods section.

We now report a more accurate description of the index that, being an absolute error, summarises the global prediction performance for the entire duration range (see P. 20, lines 11-12).

P21. Line 11-14. Please rephrase this in a more unbiased way.

The text was revised (see P. 22, line 3).

P22. Line 15. Please comment on the negative kriging weights and why these occur (screening effect, etc, see Isaaks and Srivastava, 1989 and other standard kriging references). Were there any negative weights included in the prediction results for n=6? This could lead to more extreme values of predicted TND values than the distribution of the empirical values and could then potentially lead to prediction of negative streamflow quantiles in catchments with low low-flow values, which would be good to be aware of for users of the method.

We agree with the Reviewer, we included these observations in the discussion section (by the way, negative weights were also present in the n=6 case). (P. 25, lines 15-22).

P22 line 14-15 it says that "….(2) any nested pair of catchments is associated with a high or very high Beta value." Whereas in the figure all nested catchments are on the left-hand side with small Beta values.

The Reviewer identified an error; we corrected the text, thanks (see P. 23, lines 6-8).

P22. Section 5.5.2 and Discussion. It would be interesting with more discussion about the reasons for the difference between the MAF and MAP results.*

The discussion section now reports some elements on this (see P. 27, lines 6-16).

P23. Line 3. "NSE are insensitive to n" From the figure it does not seem insensitive, NSE is getting better with a higher n (after the initial drop) with about the same difference that it is getting worse with higher n for MAF. Please rephrase this in a more unbiased way.

The text has been rephrased: "Sensitivity of NSE values to n is rather low for the study area". In fact, we do consider the sensitivity to be very limited (please, mind the range of values on the y-scale) (see P. 23, lines 20-21).

P23. Lines 18-20 is contradictory to Lines 15-17 where it says that the information was neglected after point 2 (and thus the estimation of the variogram).

We believe that there is no contradiction in the text. Point 1. of the LOOCV (leave one out cross-validation) states that empirical and theoretical variograms are computed using the entire dataset, while the description of the LNOCV (leave nested out cross-validation) clarifies that in this case also variograms are recomputed at each step. In order to make this distinction clear, we addedd a reference to point 1. in Section 5.1.1 in the revised text.

P23. Line 23. Even if the results are analogous the NSE and LNSE values could easily be reported here. It would also be interesting with some discussion as to why the performance was mainly worse for low flows when removing the nested catchments.

Low-flows are the hardest part of FDC to predict, see e.g. Figures 6, 8 and 12. Therefore, removing nested catchments has the strongest effects on this part of the curves, because there is a strong affinity in the low-flow regime as described by standardized FDC for nested catchment, definitely higher than the similarity between low-flow regimes of unnested catchments in the study area. We included this point in Section 5.5.3 (see P. 24, lines 25-28).

> *P24. Line 25-26. The fact that the prediction method is using a weighted linear average would also mean that more extreme values than in the data pool cannot be predicted (without negative kriging weights, see above).*

We included this consideration in the text (see P. 25, lines 15-22).

> *P26. Future analyses. I would suggest the following analyses as well: 1) It would be interesting to compare to other methods that use weighted combinations from dynamic pooling groups (e.g. Burn et al, 1990; Holmes et al., 2002, these could also be discussed in the introduction) and that incorporate other information than spatial proximity. 2) It would also be good to test the un-nested prediction capability against other regional models that are also applied to the same un-nested dataset.*

Thanks! We included point (1) in the text, point (2) sounds more like a specific analysis for addressing a particular case of (1) (see P. 28, lines 6-12).

> *Technical comments*
>
> *P3. Line 3. "has been" instead of "can be" when it refers to the historical period.*

Corrected, thanks!

> *P3. Line 13. "Peculiar" means "odd", "strange", "weird", etc in English and this is not a good choice of word in this type of text. This is used in several places in the text and should be replaced. In this place "…their advantage of being readily understandable…" could be used instead. On P7, line 12 "A special feature…" can be used, etc.*

Aside from what Wikipedia reports, these are the entries one can find in http://www.oxforddictionaries.com/definition/english/peculiar:

1. Different to what is normal or expected; strange
2. Particular Special

First four synonyms (form the same resource):
distinctive, characteristic, distinct, different

We revised the use of "peculiar" and "peculiarity" in the revised manuscript, but we did not feel the need to replace the adjective everywhere in the text.

> *P11. Line 25. Delete "The discriminant between the two ways resides in the fact that", it is unnecessary and start the sentence at "The uncertainty…"*

We dropped the first part. Thanks.

> *P12. Line 25-26. This sentence can be deleted since the method is presented generally for just*

"TND" in the following text.

Dropped, thanks.

P15 L18-19. Replace "each and every" with "each", here and at many other places in the manuscript to improve the language.

Changed, thanks.

P22, Line 3. 365 is a number and not "a … resampling scheme"

Corrected, thanks.

P22, line 11, end of line, add "i,j=1…18 and" after "with"

Added, thanks.

REFERENCES

Burn, D. H. 1990. "Evaluation of Regional Flood Frequency Analysis with a Region of Influence Approach." *Water Resour. Res.* 26 (10): 2257–65.

Castiglioni, S., A. Castellarin, A. Montanari, J. O. Skøien, G. Laaha, and G. Blöschl. 2011. "Smooth Regional Estimation of Low-Flow Indices: Physiographical Space Based Interpolation and Top-Kriging." *Hydrol. Earth Syst. Sc.* 15 (3): 715–27. doi:10.5194/hess-15-715-2011.

Holmes, M. G. R., A. R. Young, A. Gustard, and R. Grew. 1999. "A Region of Influence Approach to Predicting Flow Duration Curves within Ungauged Catchments." *Hydrol. Earth Syst. Sci.* 6 (4): 721–31. doi:10.5194/hess-6-721-2002.

Isaaks, E. H., and R. M. Srivastava. 1990. *Applied Geostatistics*. OUP USA.

Skøien, J. O., G. Blöschl, G. Laaha, E. Pebesma, J. Parajka, and A. Viglione. 2014. "Rtop: An R Package for Interpolation of Data with a Variable Spatial Support, with an Example from River Networks." *Computers & Geosciences* 67 (June): 180–90. doi:10.1016/j.cageo.2014.02.009.

Skøien, J. O., R. Merz, and G. Blöschl. 2006. "Top-Kriging - Geostatistics on Stream Networks." *Hydrology and Earth System Sciences* 10 (2): 277–87. doi:10.5194/hess-10-277-2006.

Manuscript prepared for Hydrol. Earth Syst. Sci. Discuss.
with version 4.1 of the LaTeX class copernicus_discussions.cls.
Date: 29 July 2014

Geostatistical prediction of flow-duration curves in an index-flow framework

A. Pugliese, A. Castellarin, and A. Brath

DICAM, Università di Bologna, Bologna, Italy

Correspondence to: A. Pugliese (alessio.pugliese3@unibo.it)

Abstract

An empirical period-of-record Flow-Duration Curve (FDC) describes the percentage of time (duration) in which a given streamflow was equaled or exceeded over an historical period of time. ~~FDCs have always attracted a great deal of interest in engineering applications because of their ability to provide a simple and yet comprehensive graphical view of the overall historical variability of streamflows in a river basin, from floods to low-flows. Nevertheless, in~~ In many practical applications one has to construct ~~FDC~~ FDCs in basins that are ungauged or where very few observations are available. We present an application strategy of ~~Topological kriging (or~~ Top-kriging~~)~~, which makes the geostatistical procedure capable of predicting ~~flow-duration curves (FDCs)~~ FDCs in ungauged catchments. Previous applications of Top-kriging mainly focused on the prediction of point streamflow indices (e.g. flood quantiles, low-flow indices, etc.)~~. In this study Top-kriging~~ ; here the procedure is used to predict ~~FDCs~~ the entire curve in ungauged sites as a weighted average of standardised empirical FDCs through the traditional linear-weighting scheme of kriging methods. ~~Our study focuses on the prediction of FDCs for 18 unregulated catchments located in Central Italy, for which daily streamflow series with length from 5 to 40 years are available, together with information on climate for the same time-span of each daily streamflow sequence. Empirical FDCs are standardised~~ In particular, we propose to standardise empirical FDCs by a reference index-flow value (i.e. mean annual flow, or mean annual precipitation times the ~~catchment~~ drainage area) and ~~the overall~~ to compute the overall negative deviation of the curves from this reference value~~is then used~~. We then propose to use these values, which we term Total Negative Deviation (TND), for expressing the hydrological similarity between catchments and for deriving the geostatistical weights. We ~~performed an extensive leave-one-out cross-validation to~~ focus on the prediction of FDCs for 18 unregulated catchments located in Central Italy, and we quantify the accuracy of the proposed technique ~~, and to compare it to traditional regionalisation models that were recently developed for the same study region~~under various operational conditions through an extensive cross-validation and sensitivity analysis. The cross-validation points out that Top-kriging is a reliable approach for predicting FDCs with Nash & Sutcliffe Efficiency measures ranging from 0.85 to 0.96~~(~~,

depending on the model settings~~) in cross-validation~~, very low biases over the entire duration range, and an enhanced representation of the low-flow regime relative to other regionalisation models that were recently developed for the same study region.

1 Introduction

An empirical Flow Duration Curve (FDC) graphically represents the percentage of time (or duration) in which the streamflow ~~can be~~ has been equalled or exceeded over a historical period of time (see e.g. Vogel and Fennessey, 1994). Empirical FDCs are often used to represent the streamflow regime of a given catchment when an adequate number of streamflow observations are available. A deterministic hydrologist would probably refer to an FDC as a key signature of the hydrological behaviour of a given basin, as it results from the interplay of climate, size, morphology, and permeability of the basin; a statistical hydrologist would refer to an FDC as the exceedance probability, or equivalently the complement to the probability distribution function (cdf) of streamflows (see e.g. Castellarin et al., 2013).

Because of their ability to provide a simple and yet comprehensive graphical view of the overall historical variability of streamflows in a river basin, from floods to low-flows, and their ~~peculiarity~~ characteristic of being readily understandable by those who do not have a strong hydrological background, empirical FDCs are routinely used in several water-related studies and engineering applications such as hydropower generation, design of water supply systems, irrigation planning and management, wasteload allocation, sedimentation studies, habitat suitability, etc. (see e.g. Vogel and Fennessey, 1995).

The literature reports two different representations of empirical flow-duration curves, depending on the reference period of time (see Vogel and Fennessey, 1994): (i) period-of-record flow duration curves (POR-FDCs), constructed on the basis of the entire observation period and (ii) annual flow duration curves (AFDCs), constructed year-wise. The two representations are complementary to each other and should be selected by practitioners depending on the water problem at hand (Castellarin et al., 2004b). For instance, AFDCs are useful for quantifying the streamflow regime in a typical hydrological year, or in a particularly wet or dry year (see Vogel

and Fennessey, 1994); POR-FDCs are a steady-state representation of the long-term streamflow regime and can be effectively used, for instance, for patching and extending streamflow data (Hughes and Smakhtin, 1996) or for assessing the long-term hydropower potential of a given site.

In many practical applications one has to predict FDCs at ungauged catchments or catchments for which the available hydrometric information is sparse (see e.g. Castellarin et al., 2013). This task is often addressed by developing regional models of FDCs. The scientific literature proposes several ~~of such modelsthat adopt~~ models, adopting different approaches to the problem: some model regard the curves as the exceedance probability function of streamflows and regionalise the parameters of theoretical frequency distributions (see Fennessey and Vogel, 1990; LeBoutillier and Waylen, 1993; Castellarin et al., 2007; Mendicino and Senatore, 2013); similarly, some other adopt a suitable mathematical expression for representing the curves and regionalise the expression parameters (Franchini and Suppo, 1996; Mendicino and Senatore, 2013); finally, some other do not make any attempt to mathematically represent the curves, they rather standardise empirical curves constructed for gauged catchments that are hydrologically similar to the target site (i.e. catchments that are characterised by a similar physiographic, pedologic and climatic conditions, also referred to as donor sites, see e.g. Kjeldsen et al., 2000) by an index streamflow (e.g. mean annual flow), and then average the dimensionless curves to predict the standardised FDC for the study catchment. The averaging procedure may (see e.g. Ganora et al., 2009), or may not (see e.g. Smakhtin et al., 1997), adopt a weighting scheme, which gives more importance to donor sites that are more hydrologically similar to the target site. The literature commonly groups these regionalisation procedures into parametric (i.e. procedures that parameterise FDCs and then regionalise parameters, like the first two examples) and non-parametric (i.e. procedure that dispense with a parameterisation of the curves, like the third example, see e.g. Castellarin et al., 2004a, 2013) procedures.

It is a common argument that an accurate representation of FDCs for daily streamflows requires probabilistic models (or mathematical expressions) with four or more parameters (LeBoutillier and Waylen, 1993; Castellarin et al., 2007), which control the position, scale and shape of the distribution. This hampers the construction of reliable regional models, due to the large

uncertainty that is commonly associated with regional relationships that express the shape parameters in terms of physiographic and climatic catchment descriptors (see Castellarin et al., 2007). As a result, Ganora et al. (2009) recently revisited the classical approach to FDCs regionalisation based on averaging standardised curves constructed for neighbouring gauged sites (Smakhtin et al., 1997), they proposed a mathematical model that enables the user to quantify the dissimilarity between empirical FDCs and associate this dissimilarity with a distance in the multidimensional space of catchment descriptors. An innovative feature of this approach is the possibility to weight each empirical FDC according to the distance between each gauged basin and the target site in the space of catchment descriptors, therefore accounting for the hydrological similarity of the donor sites with the site of interest. Like many of the traditional approaches proposed in the literature, though, the approach proposed in Ganora et al. (2009) (1) requires a preliminary subdivision of the study area into homogeneous pooling-groups of sites (i.e. clustering), (2) predicts a standardised (i.e. dimensionless) FDC for the target site, which needs then to be multiplied by a dimensional scale index (e.g. an indirect estimate of mean annual streamflow) in order to be of practical use. Both steps are critical phases of a regionalisation process. In particular concerning step (1), geostatistical regionalisation approaches have been shown to be particularly effective in dispensing with the preliminary identification of homogeneous pooling-group of sites while using regional hydrological information for predicting streamflow indices in ungauged catchments (e.g. flood quantiles, low-flow-indices, etc.: see e.g. Chokmani and Ouarda, 2004; Skøien et al., 2006; Castiglioni et al., 2009, 2011; Archfield et al., 2013; Laaha et al., 2013); yet no geostatistical procedure has been developed that specifically addresses the problem of FDC regionalisation, aside from an interpolation of the curves in the physiographic-space through a three-dimensional kriging, which is not a geostatistical procedure in the strict sense (see Castellarin, 2014).

Our paper focuses on the derivation of a geostatistical technique that addresses both limitations mentioned above for the prediction of FDCs in ungauged sites. We adopt Topological kriging or Top-kriging, which is a block-kriging with variable support area that interpolates streamflow-indices along stream networks (see e.g. Skøien et al., 2006). Top-kriging has been proved to be particularly successful in predicting point streamflow values (e.g. low-flow and

flood quantiles, mean annual flood, stream temperatures, etc.) in various geographical and climatic contexts (see e.g. Merz et al., 2008; Castiglioni et al., 2011; Vormoor et al., 2011; Archfield et al., 2013; Laaha et al., 2013).

We adopt Top-kriging as the core tool for predicting standardised (i.e. divided by mean annual flow) and dimensional long-term daily FDCs on the basis of empirical period-of-record curves (POR-FDCs, hereafter referred to as FDCs for the sake of brevity) constructed for neighbouring streamgauges.

The idea behind our study is (i) to identify a meaningful empirical point value (or index) that fully characterises the whole empirical FDC, (ii) to model the spatial correlation structure, or the spatial variability, of this index over the study region through Top-kriging and (iii) to assess the capability of this very spatial correlation model to predict FDCs in ungauged sites by weighting neighbouring empirical FDCs. We present two possible applications of the proposed procedure, the first one predicts standardised FDCs, that is FDCs divided by Mean Annual Flow (MAF), the second one predicts FDCs divided by the product between Mean Annual Precipitation (MAP) and drainage area. MAP is generally easier to predict than MAF in ungauged sites, due to the higher density of raingauging networks relative to streamgauging ones. The second application can therefore be used to predict dimensional FDCs in ungauged sites.

The approach is developed and tested through a comprehensive leave-one-out cross-validation procedure for a rather wide geographical region located in Eastern-Central Italy including 18 unregulated river basins. Castellarin et al. (2007) propose regional models of long-term daily FDCs for this area, which we use in this study as benchmark models for comparing the accuracy and reliability of the proposed approach.

2 Geostatistical hydrological prediction in ungauged sites

2.1 Top-kriging

Top-kriging is a powerful geostatistical procedure proposed by Skøien et al. (2006) which performs hydrological predictions at ungauged sites along stream-networks on the basis of the

empirical information collected at neighbouring gauging stations. As kriging techniques, the spatial interpolation is obtained in Top-kriging by a linear combination of the empirical values; therefore, the unknown value of the streamflow index of interest at prediction location x_0, $\hat{Z}(x_0)$, can be estimated as a weighted average of the variable measured in the neighborhood:

$$\hat{Z}(x_0) = \sum_{i=1}^{n} \lambda_i Z(x_i) \tag{1}$$

where λ_i is the kriging weight for the empirical value $Z(x_i)$ at location x_i, and n is the number of neighbouring stations used for interpolation. Kriging weights λ_i can be found by solving the typical ordinary kriging linear system (2), with the constrain of unbiased estimation (2b):

$$\sum_{j=1}^{n} \gamma_{i,j}\lambda_j + \theta = \gamma_{0,i} \qquad i = 1, \ldots, n \tag{2a}$$

$$\sum_{j=1}^{n} \lambda_j = 1 \tag{2b}$$

where θ is the Lagrange parameter and $\gamma_{i,j}$ is the semi-variance between catchment i and j. The semi-variance is also referred to as variogram in geostatistics and represents the space variability of the regionalised variable Z. A peculiar feature of Top-kriging is to consider the variable defined over a non-zero support S (i.e. the catchment drainage area)(Cressie, 1993; Skøien et al., 2006); this implies that the kriging system (2) remains the same, but the gamma values between the measurements need to be obtained by regularization, that is the smoothing effect of support area S on the point variogram, which is computed by applying an integral average of the variable Z over S. After this, the point variogram can be back-calculated by fitting aggregated variogram values to the sample variogram (details can be found in Skøien et al., 2006).

2.2 Total negative deviation (TND)

Top-kriging could in principle be directly applied to interpolate single streamflow values associated with a given duration (i.e. streamflow quantiles). Therefore, similarly to what proposed in Shu and Ouarda (2012), a regional prediction of FDCs could be obtained by repeatedly applying Top-krging r times, where r is the number of durations considered to provide an accurate representation of the curve (e.g. 15–20, see Shu and Ouarda, 2012), and then by interpolating the r predicted streamflow quantiles to obtain an FDC. Nevertheless, each FDC is a continuum resulting from the complex interplay between climate conditions and geomorphologic catchment characteristics (see e.g. Yaeger et al., 2012; Yokoo and Sivapalan, 2011; Beckers and Alila, 2004). This continuum would be lost, entirely or in part, by using the approach outlined above; moreover, this prediction strategy might not preserve a fundamental property of FDCs, that is the monotone (i.e. non-increasing in this paper) relationship between streamflow and duration.

Our main goal is to develop a Top-kriging procedure that regionalises the whole curve seen as a single object. In geostatistical applications one should define a "regionalised variable" to produce a characterisation of the spatial variability of the investigated phenomenon. As mentioned above, Top-kriging has been shown to be particularly reliable in predicting point (i.e. single values) streamflow indices in ungauged locations. Therefore a viable strategy could be to identify a point index that effectively summarises the entire curve, and to compute the Top-kriging λ_i values of Eq. (2) relative to this index. These values could then be used for averaging neighbouring empirical FDCs and predicting the FDC at the (ungauged) site of interest. This prediction strategy would regard each curve as a single object, and the linear interpolation of the curves (see also Sec. 3) would preserve the monotone relationship between streamflow and duration.

Some studies in the literature suggest to use the FDC slope as an overall index for the curve (see e.g. Sawicz et al., 2011). We believe though that the definition of such an index is associated with some degrees of subjectivity (e.g. which lower and upper durations to consider for the computation of the slope), and may be hard to define in some cases (e.g. ephemeral and intermittent streams).

Focusing on FDCs, Ganora et al. (2009) quantify the hydrological dissimilarity between a pair of catchments as the area between the corresponding empirical standardised (i.e. divided by mean annual flow) FDCs: two hydrologically similar catchments will show similar standardised curves, hence a small area between the curves, whereby two basins that are completely different in terms of hydrological behaviour will be characterised by highly different FDCs, and therefore the area between the curves will be large. Following this background idea, we propose to summarise the FDC through a point index which we term Total Negative Deviation (TND) between a dimensionless (i.e. standardised by a reference streamflow value) FDC and 1,

$$\mathrm{TND} = \sum_{i=1}^{m} |q_i - 1| \Delta_i \tag{3}$$

where q_i represents the i-th empirical dimensionless streamflow value, Δ_i is half of the frequency interval between the $(i+1)$-th and $(i-1)$-th streamflow values, and the summation includes only $i = 1, \ldots, m$ dimensionless streamflow values that are lower than 1 (i.e. negative deviation). m stands for the length of the dimensionless streamflow sample once values larger than 1 are excluded.

Empirical TND values are proportional to the filled areas in Fig. 1, where black thick curves represent the empirical FDCs. More specifically, Fig. 1 represents the dimensionless empirical FDCs that were constructed for three streamgauges (see Sec. 4 for a brief description of the study area) by using two standardisation methods: in one case the curve is standardised by the mean annual flow (standardisation by MAF, TND_1, top panels of Fig. 1); in the other case the curve is standardised by MAP*, that is a reference streamflow equal to the catchment area A times the mean annual precipitation MAP (standardisation by MAP*, TND_2, bottom panels in Fig. 1) (see details on standardisation procedure in Sec. 3.2).

Even though TND defined by Eq. (3) and illustrated in Fig. 1 does not describe the portion of the curve associated with low durations (high flows), it is very informative on the shape of the FDC, which, in turn, is controlled by climatic, physiographic and geo-pedological characteristics of the catchment. Catchments that are dominated by rapidly responding near-surface runoff processes have steeper FDC slopes, and therefore larger TND, while FDCs are less steep where

slower responding runoff generation processes prevail, and under these circumstances TND will be smaller. This is related to functional similarity: catchments that store and retain more water should have smaller TNDs. The magnitude of TND is related not only to the climate but also to how efficiently the catchment partitions water into runoff.

3 Top-kriging of flow-duration curves

3.1 Construction of empirical FDCs

The construction of empirical FDCs for gauged sites is straightforward: (i) pooling all observed streamflows in one sample, (ii) ranking the observed streamflows in ascending order and (iii) plotting each ordered observation vs. its corresponding duration. We adopt as duration of the i-th observation in the ordered sample in our study the estimate of the exceedance probability of the observation, $1 - F_i$. If F_i is estimated using a Weibull plotting position, the duration d_i is,

$$d_i = \text{Prob}\{Q > q_i\} = 1 - \frac{i}{N+1} \tag{4}$$

where N is the length of daily streamflows observed in a gauged site and $i = 1, \ldots, N$ is the i-th position in the rearranged sample.

A common representation of FDCs reports log-flows on the y-axis and the duration on the x-axis (see Fig. 1). Another common representation adopts a log-normal space instead, in which log-transformation of streamflows are still reported on the y-axis, while the x-axis reports duration expressed as a normal standard variate z,

$$z_i = \Phi^{-1}(1 - d_i) \tag{5}$$

where Φ is the cdf of the standard normal distribution. The combination of the two transformations improves significantly the readability of the FDC (see Fig. 2), the log-transformation enhances the representation of observed streamflows, which usually spans over two or more

orders of magnitude, while expressing the duration as a standard normal variate improves the visualization of small and large durations, that is flood- and low-flows, respectively.

3.2 Computation of empirical TND values

According to what we anticipated in Sec. 2.2, two different standardisation procedures are considered for computing TND values:

$\mathbf{TND_1}$

TND values are computed after standardisation by Mean Annual Flow (MAF), that is the traditional way to standardise FDCs.

$\mathbf{TND_2}$

TND values are computed for FDCs that are standardised by a rescaled Mean Annual Precipitation (MAP^*). The standardisation is performed by dividing each streamflow value by the empirical catchment-scale MAP value, rescaled to basin size as,

$$\mathrm{MAP}^* = \mathrm{MAP} \cdot A \cdot \mathrm{CF} \tag{6}$$

where A is the catchment area and CF is a unit-conversion factor (e.g. if streamflows are in $\mathrm{m^3\,s^{-1}}$, MAP in mm per year and A in $\mathrm{km^2}$, then ~~$\mathrm{CF} = 3.171 \times 10^{-5}\,[-]$~~ $\mathrm{CF} = 3.171 \times 10^{-5}\,[\frac{\mathrm{yr}}{\mathrm{s}} \frac{\mathrm{m^2}}{\mathrm{km^2}} \frac{\mathrm{m}}{\mathrm{mm}}]$). Once the dimensionless FDC is predicted for an ungauged site, then a dimensional FDC can be obtained by multiplying the curve by a local catchment-scale estimate of MAP^*.

The idea behind the choice of two different standardisations of FDCs derives from two different purposes: ($\mathrm{TND_1}$) MAF standardisation is the traditional choice when an index-flow regionalisation approach, with MAF being the index-flow, is used to regionalise FDCs (see Castellarin et al., 2004b; Ganora et al., 2009). Such an approach, as already mentioned, needs then an appropriate regional model for predicting the index-flow in ungauged basins (e.g. a multiregression

model) in fact, once a standardised FDC is predicted for an ungauged site, then a dimensional FDC can be obtained by multiplying the dimensionless curve by an estimate of MAF for the site of interest. Setting up a regional model for predicting MAF is a critical and delicate step in the regionalisation procedure (see e.g. Brath et al., 2001; Castellarin et al., 2004a); (TND_2) MAP* standardisation enables one to derive dimensionless FDCs to be used for regionalisation, and to predict a dimensional curve, which is ultimately what practitioners really need for addressing the water problem at hand, simply by multiplying the dimensionless FDC by MAP and catchment area. The ~~discriminant between the two ways resides in the fact that the~~ uncertainty associated with predictions of MAP is generally significantly smaller than the uncertainty associated with predictions of MAF for ungauged sites, in virtue of the large availability of rain-gauges and the accuracy of geostatistical procedure for interpolating point observations (see e.g. Brath et al., 2003; Castellarin et al., 2004a).

Concerning the practical computation of empirical TND values, that is TND_1 or TND_2, the record length generally varies among the available streamgauges. Therefore, before applying (3) one needs to set a maximum duration $d_{\max}$ that can be used in order to compute the TND values consistently for all sites in the region. $d_{\max}$ should be set according to the minimum record length in the region (e.g. if the minimum record length in the region is $5\,\mathrm{yr}$, one could set $d_{\max} = (5 \times 365)/(5 \times 365 + 1)$).

Once a suitable reference streamflow is selected for performing the standardisation of the curves (i.e. MAF or MAP*), one can easily identify the number of durations m for which the empirical dimensionless streamflow values are lower than 1 (i.e. streamflow values lower than MAF or MAP*) and compute TND according to (3). For instance, once computed the standard-normal duration z_i associated with each standardised and log-transformed streamflow quantile q_i, Δ_i in (3) can be computed as,

$$\Delta_i = 0.5\,(z_{i+1} - z_{i-1}) \qquad \text{for } i < m \tag{7a}$$

$$\Delta_i = 0.5\,(z_i - z_{i-1}) \qquad \text{for } i = 1, m. \tag{7b}$$

3.3 Geostatistical interpolation of TND and FDCs

Empirical TND (i.e. TND_1 and TND_2) values are site specific and can be interpolated with geostatistical techniques. Top-kriging can be applied as illustrated in the stepwise description by Skøien (2013) and Skøien et al. (2014) through the suite of R-functions included in the R-package `rtop`, which can be accessed from the Comprehensive R Archive Network (CRAN, http://cran.r-project.org/). The application of Top-kriging formally requires exactly the same steps in both cases (i.e. for empirical TND_1 and TND_2 values). ~~For the sake of brevity, we will recall these steps by referring to the set of empirical TND_1 values only.~~

The point sample variogram for each standardisation (see Sec. 3.2) can be computed using the binned variogram technique (see Skøien et al., 2014, for details) , for which sample points are aggregated in distance and area classes or bins, under the hypothesis of isotropy, i.e. the variogram does not vary with direction. The sample point variogram can then be modelled through a suitable theoretical model (e.g. exponential, Gaussian, spherical, fractal, etc.). Skøien et al. (2006) recommend the use of the exponential variogram.

Once the empirical variogram is modelled, the number n of neighbouring stations on which to base the spatial interpolation is set iteratively by the user on the basis of a first set of preliminary analyses, which aim at identifying the n value that produces the most accurate predictions in cross-validation (i.e. for predicting TND values in ungauged locations). This means that the local prediction of TND values, i.e. the computation of ordinary linear system in (2), depends on n-dimensional kriging weights.

We assume in our study that the n kriging weights that are computed for predicting TND in ungauged locations can also be adopted for predicting the flow-duration curve in the same locations as a weighted average of n standardised empirical curves as,

$$\hat{\psi}(x_0, d) = \sum_{i=1}^{n} \lambda_i \psi(x_i, d) \qquad d \in (0, 1) \tag{8}$$

where λ_i are the Top-kriging weights resulting from TND interpolation, $\psi(x_i, d)$ indicates the standardised empirical FDC for site x_i, that is a flow-duration curve in which streamflow quan-

tiles are divided either by MAF or by MAP*, $\hat{\psi}(x_0, d)$ stands for the standardised FDC predicted for site x_0 over the entire duration domain d, n is the number of neighbouring sites in the vicinity of the site of interest. It is worth noting that while FDC predictions are performed by using empirical standardised FDCs as a whole (i.e. the prediction is performed over the entire duration interval), the computation of empirical TND values does not consider lower durations (see details in Sec. 2.2). Therefore, it will be particularly interesting to analyse the performance of the proposed procedure for predicting high flows. We will assess our assumption relative to a study area which was extensively analysed in previous studies in the context of regionalisation of FDCs (see e.g. Castellarin et al., 2004a, 2007).

4 Study area and data

The study region includes 18 unregulated catchments, which previous studies describe as a rather heterogeneous group of sites in terms of physiographic and climatic characteristics (see e.g. Castellarin et al., 2007, 2004a). Daily streamflow series were obtained for all basins from the streamgauges belonging to the former National Hydrographic Service of Italy (SIMN) over the time period 1920–2000. The length of the observed series ranges from 5 to 40 yr (average record length: 18 yr). Also, the empirical MAP value relative to each of the 18 catchments was estimated using data collected from a rather dense raingauge network (i.e. 1 raingauge per $\approx 50\ \mathrm{km}^2$) during the same time-interval of daily streamflow observations.

Empirical FDCs were constructed from the daily streamflow series for the 18 catchments as described in Sec. 3.1. Empirical TND_1 and TND_2 values were computed for each catchment according to standardisations described in Sec. 3.2, and are illustrated in Fig. 3. As shown in the left panel of Fig. 3, empirical TND_1 values increase moving from south-east to north-west. This outcome reflects the lower perviousness of the northern catchments, which are then less capable of storing water volumes and consequently are characterised by steeper empirical FDCs. Moving from south-east to north-west, one can note for TND_2 (right panel of Fig. 3) similar patterns to those observed for TND_1 values, i.e. TND values tend to increase along the SE–NW direction. On the one hand this general behaviour suggests that in our case study Mean Annual Flow

(MAF) is largely controlled by precipitation, on the other hand, karst phenomena associated with the presence of fractured limestones result in an increase of TND_2 for the Southern catchments, i.e. sites 3006, 3003 and 3002, for which subsurface flows play a significant role.

Table 1 illustrates the variability over the study region of catchment area A (km^2), mean annual flow MAF ($m^3 s^{-1}$), mean annual precipitation MAP (mm), MAP* ($m^3 s^{-1}$), empirical TND_1 (–) and TND_2 (–) values, by reporting the minimum, mean and maximum values, together with the 1th, 2nd and 3rd quartiles of each index. For detailed information on the study area please refer to the supplementary material ("supplement" link).

5 Analysis and results

5.1 Prediction of FDCs in cross-validation

We will refer to the proposed approach as TNDTK (i.e. Total Negative Deviation Top Kriging) in the remainder of the paper. This section illustrates in detail the application of TNDTK in cross-validation, describing the accuracy of the procedure when applied in ungauged basins.

5.1.1 Standardisation by MAF

The application of TNDTK to the prediction of FDCs standardised by MAF requires the preliminary application of Top-kriging to TND_1 values, which we performed by calculating binned sample variogram first, and then by modelling binned empirical data with a 5-parameter "modified" exponential theoretical variogram (a combination of exponential and a fractal model, see details in Skøien et al., 2006). As an example, Fig. 4 ~~shows the differences between fitted variograms by either using no bins (i. e. point variogram) or by binning groups of pairs of catchments with different combinations of drainage areas. The~~ illustrates the comparison between some selected bins of the sample variogram and the regularized semivariance for those bins (see also Skøien et al., 2014, and Fig. 4 therein) . The numbers in the legend refer to different combinations of catchment areas, so that, e.g., 300 vs. 75 means the regularized variogram as a function of distance for two catchments of size ~ 300 and ~ 75 km^2, respectively;

while the solid lines represent a regularized variogram of equally sized catchments ($\sim 300\ \mathrm{km}^2$). In the same figure the black solid line represents the fitted theoretical point variogram and its five parameters were ~~fitted~~ obtained through the Weighted Least Squares (WLS) regression method from Cressie (1985) by fitting simultaneously all regularised binned variograms that were computed for various area classes (see Skøien et al., 2014) . Top-kriging was then iteratively applied to the study catchments in cross-validation to identify the most suitable number of neighbours n. Preliminary iterations indicated $n = 6$ as a good candidate for the study area (see Sec. 5.5.2).

We then used the kriging weights obtained for predicting TND_1 in cross-validation at each ~~and every~~ site to estimate dimensionless FDCs. In order to assess the prediction accuracy and to compare the performances of different models we choose to resample each curve using $p = 20$ points equally spaced in the log-normal representation (see Sec. 2.2 and Fig. 2), adopting ~~$d_1 = 0.00135$ as lower bound~~ as duration extremes $d_1 = 0.00135$ (lower bound) and $d_{20} = 0.9986$ ~~as the upper one~~ ((upper bound), where d_1 and d_{20} values are selected by referring to the minimum record length in the regional sample, i.e. 5 yr~~)~~. Predictions were performed through a weighted average, as expressed in Eq. (8), using the optimal Top-kriging cross-validation weighting scheme, i.e. λ_i with $i = 1, \ldots, n$, where $n = 6$.

As mentioned in Sec. 1, a leave-one-out cross-validation procedure (LOOCV) was performed in order to simulate ungauged conditions at each ~~and every~~ gauged site in the study area and to quantitatively test the reliability and robustness of TNDTK for predicting FDCs in ungauged basins (see examples in Kroll and Song, 2013; Salinas et al., 2013; Wan Jaafar et al., 2011; Srinivas et al., 2008).

The LOOCV ~~that~~ can be summarised by the following steps:

1. empirical and theoretical variograms are computed using the entire dataset of TND_1 values;
2. one of the gauging station, say s_i, is removed from the set of available stations;
3. a Top-kriging regional model for predicting TND_1 values is developed using the remaining $N_{\mathrm{site}} - 1$ sites;

4. TND_1 is predicted for site s_i as a weighted average of the empirical values computed for $n = 6$ neighbouring stations (see e.g. Fig. 5);

5. the weighting scheme computed in step 4 is then used to predict a standardised FDC for site s_i through Eq. (8);

6. steps from 2 to 5 are repeated $N_{\text{site}} - 1$ times.

The accuracy of the cross-validated standardised FDCs was scrupulously assessed by means of several performance indices and diagrams, which are illustrated in detail in Sec. 5.3. The algorithm described above is tailored for the proposed procedure, TNDTK, but one can implement and apply similar resampling procedures to any regional model for simulating ungauged conditions.

5.1.2 Standardisation by MAP*

Top-kriging was applied also to predict empirical TND_2 values as well as FDCs standardised by MAP*. The number of neighbouring stations n, theoretical variogram, and fitting procedure were the same as for standardisation based on MAF. We used a LOOCV analogous to the one described above (i.e. standardisation by MAF) in order to identify the weighting scheme to be used for simulating ungauged conditions for all of the study basins.

Furthermore, in order to obtain dimensional ~~prediction~~predictions, each estimated curve $\hat{\psi}(x_0, d)$ was then transformed into a dimensional FDC, as

$$\hat{\Psi}(x_0, d) = \hat{\psi}(x_0, d)\text{MAP}^*(x_0) \qquad \text{with } d \in [d_1, d_{20}] \tag{9}$$

where $\text{MAP}^*(x_0)$ indicates the local MAP* value.

5.2 Reference regional models of FDCs

The same gauged stations and data considered herein were analysed in previous studies that developed regional models of FDCs (see Castellarin et al., 2004a, 2007). This enabled us to

identify for both TNDTK applications two different reference regional models for comparing the performance of the approaches. We report here-below a brief description of such regional models.

5.2.1 Standardisation by MAF

TNDTK predictions of dimensionless FDCs were compared against the dimensionless curves predicted by two reference regional models, which we also applied in cross-validation through a LOOCV procedure:

KMOD

K model (or KMOD) is a statistical regionalisation model developed by Castellarin et al. (2007) that uses the 4-parameter unit-mean kappa distribution as parent distribution for representing standardised FDCs (see e.g. Hosking and Wallis, 1997). Three parameters, namely the parameter of location and the two shape parameters, were estimated by applying an ordinary least squares (OLS) regression algorithm. The scale parameter is derived as a function of the previous three under the hypothesis that the mean of the distribution is equal to one. Castellarin et al. (2007) regressed the parameters estimates against a suitable set of catchment descriptors through a stepwise-regression procedure in order to enable the estimation of the kappa distribution in ungauged sites. KMOD is therefore a traditional parametric regional model which we adopted as the benchmark regional model for predicting standardised FDCs (see for details Castellarin et al., 2007).

MEAN

MEAN is a simple approach to regionalisation, which neglects the physiographic and climatic heterogeneities of the study area, and predicts the standardised FDC for any ungauged site in the region as the average of all available standardised FDCs. We adopted MEAN as a baseline model due to its crude assumption and the resulting low-level accuracy.

5.2.2 Standardisation by MAP*

TNDTK predictions of dimensional FDC were compared with the predictions resulting from two benchmark models, both applied in cross-validation:

LLK

This model, based on an index-flow approach (see Castellarin et al., 2004b), adopts a two-parameter log-logistic (LL) distribution as a suitable distribution for describing the empirical frequency of the annual flow series (i.e. index-flow) and a four-parameter kappa (K) as the parent distribution for dimensionless daily streamflow frequency. Parameters of both distribution were estimated using the routine based on L-moments developed by Hosking and Wallis (see Hosking and Wallis, 1997), re-estimated through a constrained sequential quadratic programming optimisation procedure aimed at minimising the squared differences between theoretical and empirical nonexceedence probabilities, and then regressed against a suitable set of catchment descriptors through a stepwise-regression procedure. More details can be found in Castellarin et al. (2007).

KMOD

Same as KMOD for dimensionless FDCs prediction, but using a multiregression regional model to predict MAF as a function of a suitable set of catchment descriptors in ungauged basins (see e.g. Castellarin et al., 2007 for details).

5.3 Performance indices

TNDTK performance in cross-validation is analysed for both standardisation methods (MAF and MAP*) and compared with the results of reference regional models through several performance indices and diagrams. A deep analysis of model performances in terms of relative prediction residuals, i.e. relative errors between modelled and emprical values (with sign), is presented through error-duration curves. The curves show relative residuals against duration ar-

ranged in gray nested bands containing 50, 80 and 90 % of relative residuals, respectively, while a solid line illustrates the progression with duration of the median residual. Also, we use as performance descriptors the scatterdiagrams between cross-validated and empirical streamflow quantiles associated with the same duration. On the basis of the same information, NSE (Nash & Sutcliffe Efficiency) indices for each model are computed, both for natural and log-transformed streamflows. Such diagrams and indices provide a complete ~~and exhaustive~~ representation of the performance of each model in cross-validation for the entire streamflow regime, from low durations (high-flows and floods) to high ones (droughts).

Concerning the performances of the model at each site, and in particular the assessment of the number of sites for which TNDTK is more reliable than the selected reference regional models, we adopt ~~a comprehensive error index derived from the~~ an error index that summarises the prediction performance over the entire duration range by deriving the distance between predicted and empirical FDCs, as proposed in Ganora et al. (2009):

$$\delta_{\text{mod}} = \sum_{k=1}^{p} |q_{k,\text{emp}} - \hat{q}_{k,\text{mod}}| \tag{10}$$

where $p = 20$ resampled points, while $q_{k,\text{emp}}$ and $\hat{q}_{k,\text{mod}}$ stand for the empirical and predicted streamflow quantiles (dimensionless or dimensional, depending on the application) ranked at the kth duration.

5.4 Results

5.4.1 Standardisation by MAF: dimensionless FDCs

Figure 5 (left) reports empirical TND_1 values against their Top-kriging predictions in cross-validation. The overall NSE is 0.81. In the same figure one can observe a poor prediction (i.e. significant underprediction) for site 3701, which can be interpreted as a result of the very high empirical TND value obtained for that site (site 3701, $\text{TND}_1 = 9.8\,[-]$, $A = 605\,[\text{km}^2]$), the largest in the study region.

Concerning the predictions of standardised FDCs, the error-duration curves of Fig. 6 clearly shows that TNDTK significantly outperforms KMOD and MEAN: the distribution of relative residuals plotted against duration is characterised by narrower bands (50, 80 and 90 % of the relative errors) for the entire duration interval, even though this behaviour is more marked for lower durations. The progression with duration of the median residual (black thick line) in the same figure highlights unbiasedness being close to zero for the entire duration interval. Scatterdiagrams between predicted and observed standardised flows indicate high accuracy of TNDTK, with NSE $= 0.958$ and LNSE $\simeq 0.96$, the latter computed for log-flows. MEAN and KMOD are associated with lower NSE and LNSE values.

Finally, Fig. 7 presents the overall absolute error for each site. In particular in Fig. 7 scatterdiagrams of δ_{mod} are illustrated in two panels, where the x-axes reports errors computed for the proposed model (TNDTK) while the y-axes reports in turn errors from reference models. In this representation an equivalence between model performances is represented by the solid bisecting line; hence if one point falls in the top-left above the 1:1-line TNDTK provides better predictions than the reference model, otherwise if it falls below the 1:1-line. Figure 7 clearly shows that KMOD is less accurate than TNDTK for 14 out of 18 sites, while MEAN performs the poorest, with 16 out of 18 sites characterised by higher δ values relative to TNDTK.

5.4.2 Standardisation by MAP*: dimensional FDCs

Right panel of Fig. 5 highlights satisfactory performance of Top-kriging for predicting TND_2 values in ungauged basins, NSE value is approximately 0.6, and site 3701 still presents an outlying behaviour for the same reason explained before.

Although the cross-validated predictions of TND_2 are less accurate than TND_1, TNDTK performance for predicting dimensional FDCs is good. Comparing TNDTK with LLK models, Fig. 8 shows for LLK narrower bands for $d < 0.8$, particularly the band illustrating 90 % of residuals, while in the low-flow range (i.e. $0.8 < d < 1$) TNDTK shows slightly better performances, resulting in narrower error bands. The bottom panels in the same figure report the scatterdiagrams of predicted vs. observed dimensional flows, expressing the goodness and reliability of TNDTK when used for predicting dimensional FDC on the basis of MAP. Even

though TNDTK shows an NSE $= 0.914$, which is lower than the NSE value associated with LLK and equal to KMOD one, TNDTK is associated with the highest LNSE value (i.e. 0.922), ~~which highlights the very good performance~~ highlighting a good performance of TNDTK for low-flows. Figure 9 confirms good performance of TNDTK against LLK and KMOD, showing in both cases better accuracy for 10 out of 18 catchments. Also, among the 8 catchments for which LLK and KMOD perform better than TNDTK, it is worth nothing that performances are practically coincident with TNDTK in 2 cases for LLK (i.e. sites 3006 and 2201) and 3 cases for KMOD (i.e. sites 1004, 2101 and 3006).

5.5 Sensitivity analysis

5.5.1 Consistency of the kriging weighting scheme

The core assumption of the proposed method is that Top-kriging weights λs identified for predicting TND values can be used to weight empirical FDCs. In order to test and validate this assumption we analysed the relationship between such weights and the degree of dissimilarity between empirical FDCs. In particular, we computed for each pair of catchments a dissimilarity metric between catchment i and j, $\beta_{i,j}$, proposed by Ganora et al. (2009), which can be expressed as follows~~for catchement i and j~~:

$$\beta_{i,j} = \sum_{k=1}^{365} |q_{i,k} - q_{j,k}| \tag{11}$$

where 365 is ~~a reasonable resampling scheme~~ the number of points used for the resampling and $q_{i,k}$ and $q_{j,k}$ are the streamflow values associated with duration $d_k = \frac{k}{365+1}$ for site i and j respectively. If our assumption is correct, large β values (i.e. dissimilar curves) should be associated with small λ values, and vice-versa. Top-kriging takes into account the nested structure of catchments, therefore where the upstream-downstream correlation occurs (i.e. similar curve with small β) relative high λ value is expected.

Figure 11 (right panel) plots $\beta_{i,j}$ values computed with Eq. (11) for each pair of basins in the study area, with $i, j = 1, \ldots, 18$ and $i \neq j$ (i.e. 306 points), against the corresponding $\lambda_{i,j}$

weights obtained by running a TNDTK session with TND = TND_1 and, necessarily, a number of neighbours $n = 17$ (i.e. all stations need to be considered if we have to compare $\beta_{i,j}$ with $\lambda_{i,j}$ for $i, j = 1, \ldots, 18$ and $i \neq j$). The figure also highlights the differences between nested (large black dots) and un-nested (gray circles) catchments pairs. The figure clearly proves that the hypotheses are satisfied: (1) weights $\lambda_{i,j}$ show a descending pattern as $\beta_{i,j}$ values increase and (2) ~~any~~ none of the nested pair of catchments ~~is~~ presents kriging weight λ associated with a high or very high β value ~~.~~ (i.e. all nested catchments are on the left-hand side with small β values).

5.5.2 Sensitivity to the number of neighbours n

As mentioned in Sec. 5.1.1 and 5.1.2, we set the number of neighbours $n = 6$ in Eq. (8) for performing the prediction of FDCs. We identified this value through a sensitivity analysis, which was carried out by running multiple Top-kriging sessions, each one referring to a different n value. The main outcome of our sensitivity analysis is that the performance of the approach is not dramatically dependent on n, quite the opposite. Figure 10 shows the results of the sensitivity analysis for both standardisations (i.e. MAF and MAP^*) obtained in each session in terms of NSE and LNSE for n, ranging from 3 to 17 (i.e. being 18 the total number of catchments for the study area). The left panel refers to dimensionless FDCs (i.e standardisation by MAF) and shows for $n = 6$ the best trade-off between NSE and LNSE. Nevertheless, NSE and LNSE are rather high for all n values. Likewise, the right panel refers to the prediction of dimensional FDCs (i.e. standardisation by MAP^*) and it shows that performances in ~~termes of NSE are insensitive~~ terms of sensitivity of NSE values to n is rather low for the study area, while in terms of LNSE, we obtain slightly better performences are associated with $n \leq 6$. As a result of the analysis we selected $n = 6$ for all applications for the sake of consistency, even though selecting a different value for n does not impact the results significantly.

5.5.3 Sensitivity to the degree of nesting of the study catchments

From an operational view point it is important to understand if the degree of nesting of the study catchments impacts the performance of the approach. Better performances are to be expected in all those cases in which empirical FDCs can be constructed upstream or downstream the (ungauged) site of interest. In order to quantify this impact we validated TNDTK by removing all catchments that are nested with the catchment of interest. Figure 11 (left panel) shows all nested pairs through a graphical matrix where nested pairs are highlighted with large black dots (catchment IDs are also indicated). First we identified all nested pairs of catchments (i.e. basin-subbasin relationships). Second, we used a cross-validation procedure similar to the procedure described in Sec. 5.1.1, in which, at point 2, we neglected all information collected for the site of interest, but also upstream or downstream that site. We termed this procedure Leave Nested Out Cross-Validation (LNOCV). It is worth noting that LNOCV estimates empirical and theoretical variograms at each ~~and every~~ step of the validation procedure, differently from LOOCV, where they are estimated beforehand once and for all ~~.~~ (see point 1. in Sec. 5.1.1).

We report here only the results referring to the prediction of dimensionless FDCs (i.e. standardisation by MAF). Results obtained relative to dimensional FDCs (i.e. standardisation by MAP*) are analogous. The results, shown in Fig. 12, highlight a slight reduction of performances, with NSE and LNSE indices equal to 0.95 and 0.92 respectively (central panel)~~; also~~. In particular, looking at the error-duration bands (left panel in the same figure) the distribution of relative residuals presents slightly wider bands and a lager bias for the median line, especially relative to the high durations (low flows). Moreover, comparing the overall error index for each site produced by the two cross-validations (i.e. LOOCV and LNOCV) (right panel), most of the points (14 out of 18) falls above the solid bisecting line, confirming an impoverished prediction capability of the latest approach. Nevertheless, the detriment of performances ~~associated~~ obtained with LNOCV appears to be limited ~~.~~ and associated in particular with the low-flow regime (high duration values). This was to be expected as this portion of FDC is the hardest to predict (see e.g. Figures 6, 8 and 12 and Castellarin et al., 2004a) , and therefore not considering catchments having their outlet located upstream or downstream the target site has

the strongest effects due to the strong hydrological affinity of these catchments with the target one (i.e. they share local climate as well as physiographic and geological characteristics, see e.g. Laah

6 Discussion and future work

6.1 Is Top-kriging suitable for predicting long-term FDCs?

The results of the cross-validation show that Top-kriging can be effectively applied for predicting standardised FDCs (i.e. flow-duration curves divided by the mean annual flow, MAF) in the study region. In particular, the interpolation strategy applied in this study (termed Total Negative Deviation Top-kriging, TNDTK), that is (1) the computation of the streamflow index Total Negative Deviation (TND) for empirical standardised FDCs, (2) the modelling of spatial correlation of empirical TND values along the stream network, (3) the identification of a linear weighting scheme for averaging empirical dimensionless FDCs on the basis of the correlation model identified at step (2), results in reliable predictions of standardised FDCs in ungauged sites.

It is worth highlighting that the application of the procedure may produce negative weights (see Fig. 11 for the case in which the number of neighbours in set to $n = 17$). Negative weights are often the result of the so-called screening effect
(i.e. remote data points are screened by set of closer data locations in front of them, see e.g. Deutsch, which can be accentuated by a zero-nugget variogram model, as our this case. We did not experience adverse effects associated with negative weights in our analysis, but, in case the presence of negative weights results in non-physical estimates (e.g. negative streamflow values) one may set all weights to be positive trough the rtop routine options (see Skøien et al., 2014) .

The curves predicted in cross-validation are unbiased for the entire duration range (i.e. from high- to low-flows) and the prediction residuals are as small as, or smaller than, the residuals resulting from the application of traditional regionalisation schemes. Analysing the results in detail, Fig. 7 indicates that TNDTK performed significantly worse than the baseline and bench-

mark regional models in three cases only. The benchmark model (i.e. KMOD) better predicts the FDC for site 3701 (left panel of Fig. 7). As illustrated in right panel in Fig. 2, site 3701 is associated with the steepest empirical flow duration curve of the study region and therefore the highest empirical TND value (see Table 1 and Figs. 1 and 5).

The core assumption of Top-kriging hypothesises is that hydrological similarity is mainly controlled by spatial proximity, and this may represent an important limitation in some regions where geology and/or morphology have a large impact on streamflows, such that the hydrological regime of nearby catchments may be quite different. This could in principle explain the poor prediction obtained in the study for site 3701, which is characterised by a very limited permeability (i.e. can be regarded as impervious) relative to the surrounding catchments, and, consequently, a much steeper empirical FDC than the neighbouring sites. Conversely, information on permeability is explicitly incorporated in the multiregression models included in KMOD (see e.g. Castellarin et al., 2007). Furthermore, the baseline model MEAN significantly outperforms TNDTK for sites 2502 and 801, and this result can be explained by noting that both sites are associated with empirical standardised curves that are well represented by the average standardised FDC for the study region (see right panel in Fig. 2 and Castellarin, 2014), that is the curve associated with the baseline regional model (MEAN) in cross-validation.

Aside from peculiar cases highlighted above, TNDTK shows a high performance in cross-validation that is likely to result from several advantages of the proposed procedure. TNDTK dispenses with the critical phase of delineating hydrologically homogeneous pooling group of sites (see Castellarin et al., 2004a) by exploiting the spatial correlation structure of the streamflow regime (see Archfield and Vogel, 2010). Nevertheless, the approach does not require to set up multiregression models for estimating the parameters of a mathematical expression (e.g. a theoretical frequency distribution) controlling the shape of the curve, which are often associated with a large uncertainty and limited robustness (see Castellarin et al., 2007); TNDTK predicts the shape of the curve for an ungauged basin through a non-parametric procedure as a weighted average of empirical standardised FDCs (e. g. Smakhtin et al., 1997; Ganora et al., 2009). The weighting scheme also ensures for the predicted curve a non–increasing (i.e. mono-

tone) relationship between streamflow and duration, which is one of the main properties of flow-duration curves.

The study also points out that TNDTK can be used for predicting dimensional FDCs in ungauged sites on the basis of a minimal set of hydrological information, that is (a) empirical FDCs for a group of gauged basins and (b) an estimate of Mean Annual Precipitation (MAP) for all gauged basins in the region, as well as for the target ungauged basin. By comparing Figures 6 and 7 with Figures 8 and 9 one may get the impression that a standardization of streamflows by MAP* reduces TNDTK performance relative to a standardizations by MAF. It is worth pointing out that one cannot directly compare the results for these two cases since Figures 6 and 7 (standardization by MAF) refer to the prediction of a dimensionless FDCs, while Figures 8 and 9 (standardization by MAP*) refer to the prediction of dimensional FDCs. Moreover, concerning the prediction of dimensional FDCs (standardization by MAP*), the similar performances between TNDTK and the benchmark regional models are rather surprising; while the benchmark regional models incorporate a regionalization of empirical mean annual flows, TNDTK uses only local information on precipitation for predicting a dimensional FDC in the target site. Even though TNDTK does not show a clear supremacy relative to more traditional approaches~~(see Figs. 8 and 9)~~, it has to be highlighted that its application is rather straightforward and does not require any subjective choice, which, together with the fact that the procedure can be implemented with a limited amount of input data, makes TNDTK a very interesting alternative for predicting dimensional FDCs.

6.2 Future analyses

Our study is evidently a preliminary analysis, which tackles the exploration of geostatistical approaches for predicting FDCs. Therefore, the results of our study open up several possible research avenues. In particular, we focus on the prediction of long-term steady-state FDCs, on the basis of Period-of-Record (POR) empirical FDCs. Applicability of TDNTK to the prediction of annual FDCs for typical hydrologic years, as well as for particularly wet or dry years (see e.g. Vogel and Fennessey, 1994; Castellarin et al., 2004b), is an open problem that needs to be specifically and quantitatively addressed. Evidently, the proposed approach needs to be

further investigated in other geographical contexts. In particular, the application of TNDTK for predicting dimensional FDCs on the basis of catchment-scale MAP values deserves some ~~further~~ additional tests that aim at verifying its suitability for significantly different climatic conditions (e.g. arid regions, alpine catchments, etc.), in which the streamflow regime is not heavily controlled by the rainfall regime, as for the considered case study.

Also, future analyses will focus on a comparison between TNDTK with other methods that use weighted combinations from dynamic pooling-groups of sites, such as the Region of Influence (RoI) approach (e.g. Burn, 1990; Holmes et al., 2002) . This will enable a better understanding of the potential of geostatistical techniques and the informativeness of spatial structure of signatures of the streamflow-regime, such as TND, relative to approaches that incorporate other information than spatial proximity when it comes to the prediction of FDCs in ungauged sites (see e.g. Merz and Blöschl, 2005, for the prediction of flood quantiles) .

Finally, we propose to summarise empirical flow-duration curves through the index TND, which expresses the total negative deviation of the curve from a reference streamflow value. We are aware that the proposed procedure needs to be further tested in different geographical and climatic contexts before its general validity can be acknowledged. Also, we believe that the TND index identified in this study incorporates a worth of hydrological information and has the potential to be extremely useful in a number of hydrological problems other than the prediction of FDCs, such as catchment classification (see Wagener et al., 2007; Di Prinzio et al., 2011) or regionalisation studies (Laaha and Blöschl, 2006; Gaál et al., 2012). Future analyses will specifically address these points. Moreover, future analyses ~~should~~ will focus on the identification of a global indicator of the similarity between FDCs to be used to analyse and model geographical correlation between the empirical curves themselves, this would enable one to base the definition of the linear weighting scheme on a more comprehensive and descriptive indicator of the streamflow regime.

7 Conclusions

This study explores the possibility to extend the application of Top-kriging, which is generally used for spatial interpolation of point streamflow indices (e.g., estimated flood quantiles, low-flow indices, temperature, etc.), to the prediction of period-of-record flow-duration curves (FDCs) in ungauged basins. Top-kriging is used in this study to geostatistically interpolate standardised FDCs along the stream network of a broad geographical area in Central-Eastern Italy. We identify the linear weighting typical of any kriging procedure by modeling the spatial correlation structure of an empirical streamflow index, which was shown in the study to be particularly useful in describing the daily streamflow regime of a given catchment. In particular, we define the index, which we term Total Negative Deviation (TND), as the overall negative deviation of an empirical FDC relative to a reference streamflow-value used for the standardisation of the curve itself. We consider two different reference streamflow values, that is the Mean Annual Flow (MAF) and catchment-scale Mean Annual Precipitation times the drainage area of the catchment (MAP^*), and we use these streamflow values for standardisation of the empirical FDCs prior to regionalisation. The standardisation based on MAF enables us to develop a Top-kriging-based regional model of dimensionless FDCs, while the standardisation based on MAP^* enables us to predict dimensional flow-duration curves in ungauged basins via Top-kriging. The two regional estimators were cross-validated and compared in terms of prediction performances with other regional models of dimensionless and dimensional flow-duration curves that were previously developed for the study area. The comparison highlights good performances of the proposed procedure, which we termed Total Negative Deviation Top-kriging (TNDTK) relative to traditional regional models. TNDTK is unbiased throughout the entire duration interval and characterised by particularly small residuals for high durations (i.e. improved predictions of low-flows). Moreover, the prediction accuracy of TNDTK is similar to, or higher than, more complex regionalisation approaches that use multiregression models incorporating information on the permeability, morphology, climate, etc. of the catchment. This result seems to confirm the value of spatial proximity relative to catchment attributes (see e.g. Merz and Blöschl, 2005) when hydrological predictions in ungauged basins are concerned.

Acknowledgements. We thankfully acknowledge Jon O. Skøien for his helpful assistance with Top-kriging applications via `rtop` R-package. We also ~~thank referee~~ would like to thank Daniele Ganora and ~~Anonymous referee 2 for their valuable advices, enhancing the scientific contribution of this manuscript.~~ an anonymous reviewer for their suggestions and valuable help in improving the overall quality of this paper. The study is part of the research activities carried out by the working group: Anthropogenic and Climatic Controls on WateR AvailabilitY (ACCuRAcY) of Panta Rhei - Everything Flows Change in Hydrology and Society (IAHS Scientific Decade 2013-2022).

References

Archfield, S. and Vogel, R.: Map correlation method: Selection of a reference streamgage to estimate daily streamflow at ungaged catchments, Water Resour Res, 46, doi: 10.1029/2009WR008481, 2010.

Archfield, S. A., Pugliese, A., Castellarin, A., Skøien, J. O., and Kiang, J. E.: Topological and canonical kriging for design flood prediction in ungauged catchments: an improvement over a traditional regional regression approach?, Hydrology and Earth System Sciences, 17, 1575–1588, doi: 10.5194/hess-17-1575-2013, 2013.

Beckers, J. and Alila, Y.: A model of rapid preferential hillslope runoff contributions to peak flow generation in a temperate rain forest watershed, Water Resour Res, 40, doi: 10.1029/2003WR002582, 2004.

Blöschl, G., Sivapalan, M., Thorsten, W., Viglione, A., and Savenije, H.: Runoff prediction in ungauged basins: synthesis across processes, places and scales, Cambridge University Press, ISBN: 9781107028180, 2013.

Brath, A., Castellarin, A., Franchini, M., and Galeati, G.: Estimating the index flood using indirect methods, Hydrol Sci J, 46, 399–418, doi: 10.1080/02626660109492835, 2001.

Brath, A., Castellarin, A., and Montanari, A.: Assessing the reliability of regional depth-duration-frequency equations for gaged and ungaged sites, Water Resour. Res., 39, doi: 10.1029/2003WR002399, 2003.

Burn, D. H.: Evaluation of regional flood frequency analysis with a region of influence approach, Water Resour Res, 26, 2257–2265, 1990.

Castellarin, A.: Regional Prediction of Flow-Duration Curves Using a Three-dimensional Kriging, J Hydrol, doi: 10.1016/j.jhydrol.2014.03.050, 2014.

Castellarin, A., Galeati, G., Brandimarte, L., Montanari, A., and Brath, A.: Regional flow-duration curves: reliability for ungauged basins, Adv Water Resour, 27, 953–965, doi: 10.1016/j.advwatres.2004.08.005, 2004a.

Castellarin, A., Vogel, R., and Brath, A.: A stochastic index flow model of flow duration curves, Water Resour Res., 40, W03 104, doi: 10.1029/2003WR002524, 2004b.

Castellarin, A., Camorani, G., and Brath, A.: Predicting annual and long-term flow-duration curves in ungauged basins, Adv Water Resour, 30, 937–953, doi: 10.1016/j.advwatres.2006.08.006, 2007.

Castellarin, A., Botter, G., Hughes, D. A., Ouarda, T. B. M. J., and Parajka, J.: Prediction of flow duration curves in ungauged basins, chap. 7, pp. 135–162, Runoff prediction in ungauged basins: synthesis across processes, places and scales, Cambridge University Press, ISBN: 9781107028180, 2013.

Castiglioni, S., Castellarin, A., and Montanari, A.: Prediction of low-flow indices in ungauged basins through physiographical space-based interpolation, J Hydrol, 378, 272–280, doi: 10.1016/j.jhydrol.2009.09.032, 2009.

Castiglioni, S., Castellarin, A., Montanari, A., Skøien, J. O., Laaha, G., and Blöschl, G.: Smooth regional estimation of low-flow indices: physiographical space based interpolation and top-kriging, Hydrol and Earth Sys Sci, 15, 715–727, doi: 10.5194/hess-15-715-2011, 2011.

Chokmani, K. and Ouarda, T. B. M. J.: Physiographical space-based kriging for regional flood frequency estimation at ungauged sites, Water Resources Research, 40, W12 514, doi: 10.1029/2003WR002983, 2004.

Cressie, N.: Fitting variogram models by weighted least squares, J Int Ass Math Geol, 17, 563–586, doi: 10.1007/BF01032109, 1985.

Cressie, N. A. C.: Statistics for spatial data, Wiley series in probability and mathematical statistics: Applied probability and statistics, J. Wiley, ISBN: 9780471002550, 1993.

Dalrymple, T.: Flood-frequency analyses, Manual of Hydrology: Part 3, Tech. Rep. WSP - 1543-A, United States Geological Survey, 1960.

Deutsch, C. V.: Correcting for negative weights in ordinary kriging, Comput Geoscie, 22, 765 – 773, doi:10.1016/0098-3004(96)00005-2, 1996.

Di Prinzio, M., Castellarin, A., and Toth, E.: Data-driven catchment classification: application to the pub problem, Hydrol. Earth Syst. Sci., pp. 1921–1935, doi: 10.5194/hess-15-1921-2011, 2011.

Fennessey, N. and Vogel, R.: Regional Flow-Duration Curves for Ungauged Sites in Massachusetts, J Water Res Pl-ASCE, 116, 530–549, doi: 10.1061/(ASCE)0733-9496(1990)116:4(530), 1990.

Franchini, M. and Suppo, M.: Regional Analysis of Flow Duration Curves for a Limestone Region, Water Resour Manag, 10, 199–218, doi: 10.1007/BF00424203, 1996.

Gaál, L., Szolgay, J., Kohnová, S., Parajka, J., Merz, R., Viglione, A., and Blöschl, G.: Flood timescales: Understanding the interplay of climate and catchment processes through comparative hydrology, Water Resources Research, p. W04511, doi: 10.1029/2011WR011509, 2012.

Ganora, D., Claps, P., Laio, F., and Viglione, A.: An approach to estimate nonparametric flow duration curves in ungauged basins, Water Resour Res, 45, doi: 10.1029/2008WR007472, 2009.

Holmes, M., Young, A., Gustard, A., and Grew, R.: A region of influence approach to predicting flow duration curves within ungauged catchments, Hydrol Earth Sys Sci, 6, 721–731, 2002.

Hosking, J. R. M. and Wallis, J. R.: Regional Frequency Analysis: An Approach Based on L-Moments, Cambridge University Press, ISBN: 9780521019408, 1997.

Hughes, D. A. and Smakhtin, V.: Daily flow time series patching or extension: A spatial interpolation approach based on flow duration curves, Hydrol Sci J, 41, 851–871, doi: 10.1080/02626669609491555, 1996.

Isaaks, E. H. and Srivastava, R. M.: Applied Geostatistics, Oxford University Press, USA, 1990.

Kjeldsen, T. R., Lundorf, A., and Rosbjerg, D.: Use of a two-component exponential distribution in partial duration modelling of hydrological droughts in Zimbabwean rivers, Hydrol Sci J, 45, 285–298, doi: 10.1080/02626660009492325, 2000.

Kjeldsen, T. R., Smithers, J. C., and Schulze, R. E.: Regional flood frequency analysis in the KwaZulu-Natal province, South Africa, using the index-flood method, J Hydrol, 255, 194–211, doi: 10.1016/S0022-1694(01)00520-0, 2002.

Kroll, C. N. and Song, P.: Impact of multicollinearity on small sample hydrologic regression models, Water Resour. Res., 49, 3756–3769, doi: 10.1002/wrcr.20315, 2013.

Laaha, G. and Blöschl, G.: A comparison of low flow regionalisation methods - catchment grouping, J Hydrol, 323, 193–214, 2006.

Laaha, G., Skøien, J. O., Nobilis, F., and Blöschl, G.: Spatial Prediction of Stream Temperatures Using Top-Kriging with an External Drift, Environ Model Assess, pp. 1–13, doi: 10.1007/s10666-013-9373-3, 2013.

Laaha, G., Skøien, J., and Blöschl, G.: Spatial prediction on river networks: comparison of top-kriging with regional regression, Hydrological Processes, 28, 315–324, doi:10.1002/hyp.9578, 2014.

LeBoutillier, D. W. and Waylen, P. R.: A stochastic model of flow duration curves, Water Resour Res, 29, 3535–3541, doi: 10.1029/93WR01409, 1993.

Mendicino, G. and Senatore, A.: Evaluation of parametric and statistical approaches for the regionalization of flow duration curves in intermittent regimes, J Hydrol, 480, 19–32, doi: 10.1016/j.jhydrol.2012.12.017, 2013.

Merz, R. and Blöschl: Flood frequency regionalisation - spatial proximity vs. catchment attributes, J Hydrol, pp. 283–306, 2005.

Merz, R., Blöschl, G., and Humer, G.: National flood discharge mapping in Austria, Nat Hazards, 46, 53–72, doi: 10.1007/s11069-007-9181-7, 2008.

Niadas, I. A.: Regional flow duration curve estimation in small ungauged catchments using instantaneous flow measurements and a censored data approach, J Hydrol, 314, 48–66, doi: 10.1016/j.jhydrol.2005.03.009, 2005.

Salinas, J. L., Laaha, G., Rogger, M., Parajka, J., Viglione, A., Sivapalan, M., and Blöschl, G.: Comparative assessment of predictions in ungauged basins – Part 2: Flood and low flow studies, Hydrol. and Earth Syst. Sci., 17, 2637–2652, doi: 10.5194/hess-17-2637-2013, 2013.

Sawicz, K., Wagener, T., Sivapalan, M., Troch, P. A., and Carrillo, G.: Catchment classification: empirical analysis of hydrologic similarity based on catchment function in the eastern USA, Hydrology and Earth System Sciences Discussions, 8, 4495–4534, doi: 10.5194/hessd-8-4495-2011, 2011.

Shu, C. and Ouarda, T. B. M. J.: Improved methods for daily streamflow estimates at ungauged sites, Water Resour Res, 48, doi: 10.1029/2011WR011501, 2012.

Skøien, J., Blöschl, G., Laaha, G., Pebesma, E., Parajka, J., and Viglione, A.: rtop: An R package for interpolation of data with a variable spatial support, with an example from river networks, Comput Geosci, 67, 180 – 190, doi: 10.1016/j.cageo.2014.02.009, 2014.

Skøien, J. O.: rtop: Interpolation of data with variable spatial support, r package version 0.3-45, 2013.

Skøien, J. O., Merz, R., and Blöschl, G.: Top-kriging - geostatistics on stream networks, Hydrol Earth Syst Sci, 10, 277–287, doi: 10.5194/hess-10-277-2006, 2006.

Smakhtin, V. Y., Hughes, D. A., and Creuse-Naudin, E.: Regionalization of daily flow characteristics in part of the Eastern Cape, South Africa, Hydrol Sci J, 42, 919–936, doi: 10.1080/02626669709492088, 1997.

Srinivas, V., Tripathi, S., Rao, A. R., and Govindaraju, R. S.: Regional flood frequency analysis by combining self-organizing feature map and fuzzy clustering, J. of Hydrol., 348, 148–166, doi: 10.1016/j.jhydrol.2007.09.046, 2008.

Vogel, R. M. and Fennessey, N. M.: Flow-Duration Curves. I: New Interpretation and Confidence Intervals, J Water Res Pl - ASCE, 120, 485–504, doi: 10.1061/(ASCE)0733-9496(1994)120:4(485), 1994.

Vogel, R. M. and Fennessey, N. M.: Flow Duration Curves II: A Review of Applications in Water Resources Planning, J Am Water Resour As, 31, 1029–1039, doi: 10.1111/j.1752-1688.1995.tb03419.x, 1995.

Vormoor, K., Skaugen, T., Langsholt, E., Diekkrüger, B., and Skøien, J. O.: Geostatistical regionalization of daily runoff forecasts in Norway, Intl J River Basin Manag, 9, 3–15, doi: 10.1080/15715124.2010.543905, 2011.

Wagener, T., Sivapalan, M., Troch, P., and Woods, R.: Catchment classification and hyrdrolig similarity, Geography compass, 1/4, 901–931, doi: 10.1111/j.1749-8198.2007.00039.x, 2007.

Wan Jaafar, W. Z., Liu, J., and Han, D.: Input variable selection for median flood regionalization, Water Resour. Res., 47, W07 503, doi: 10.1029/2011WR010436, 2011.

Yaeger, M., Coopersmith, E., Ye, S., Cheng, L., Viglione, A., and Sivapalan, M.: Exploring the physical controls of regional patterns of flow duration curves – Part 4: A synthesis of empirical analysis, process modeling and catchment classification, Hydrology and Earth System Sciences, 16, 4483–4498, doi: 10.5194/hess-16-4483-2012, 2012.

Yokoo, Y. and Sivapalan, M.: Towards reconstruction of the flow duration curve: development of a conceptual framework with a physical basis, Hydrology and Earth System Sciences, 15, 2805–2819, doi: 10.5194/hess-15-2805-2011, 2011.

Table 1. Study catchments: variability of drainage area (A), Mean Annual Flow (MAF), Mean Annual Precipitation (MAP), rescaled mean annual precipitation (MAP^*), empirical TND_1 and TND_2 and length of the observed streamflow series (Y); minimum, maximum, mean, 1st, 2nd (median) and 3rd quartiles of the sample distributions.

	A $[km^2]$	MAF $[m^3 s^{-1}]$	MAP $[mm]$	MAP^* $[m^3 s^{-1}]$	TND_1 [–]	TND_2 [–]	Y [yr]
min	61	1.49	918	2.17	1.59	1.25	5
1st Qu.	104	2.63	1079	3.60	2.76	4.38	8.5
median	164	3.83	1123	5.99	3.82	5.78	11.5
mean	330	6.51	1118	11.69	4.52	6.11	18.1
3rd Qu.	562	7.54	1162	17.53	5.74	7.55	26
max	1044	21.29	1298	37.07	9.83	13.21	40

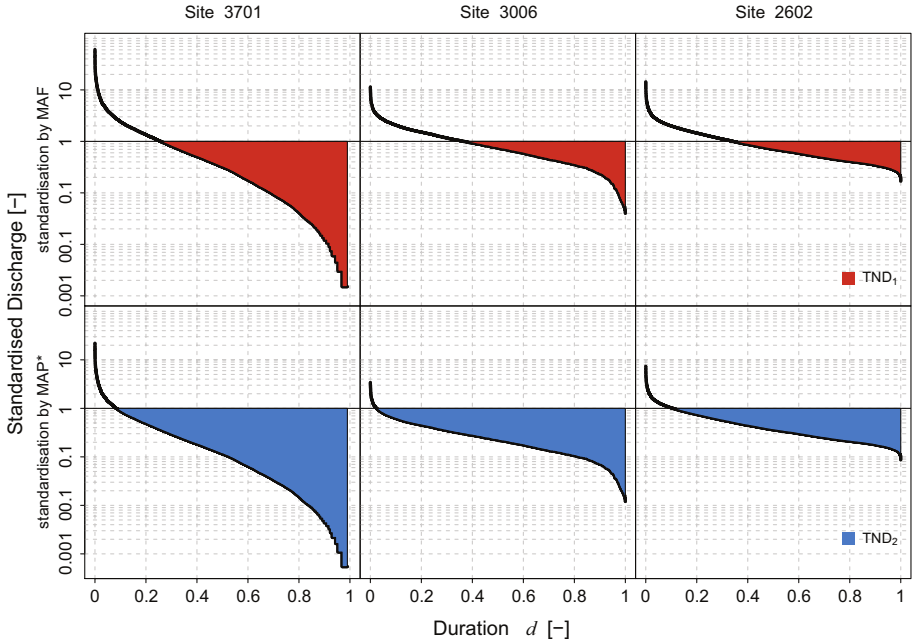

Fig. 1. Total Negative Deviation (TND, filled area) for three catchments with different hydrological behaviours (see Sec. 4). Top panels: TND_1 (red area) for an empirical FDC (black thick line) standardised by Mean Annual Flow (MAF); bottom panels: TND_2 (blue area) for an empirical FDC (black tick line) standardised by $MAP^* = MAP \cdot A \cdot CF$, where MAP is the Mean Annual Precipitation, A is the drainage area and CF is a unit-conversion factor.

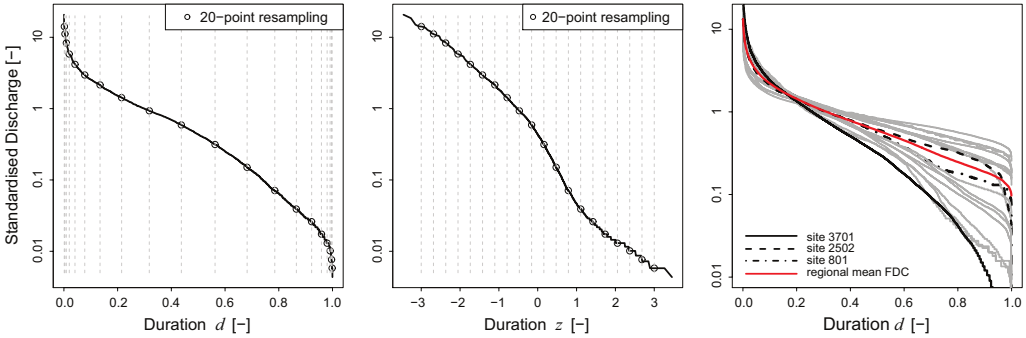

Fig. 2. FDC representations: log-natural scale (left), log-normal scale (center); the panels also show a resampling of the empirical curve (circles) which employs 20 equally-spaced points in the standard-normal space; standardised empirical FDCs for the study region (right), FDC for sites 3701, 801, 2502 and regional mean FDC are highlighted.

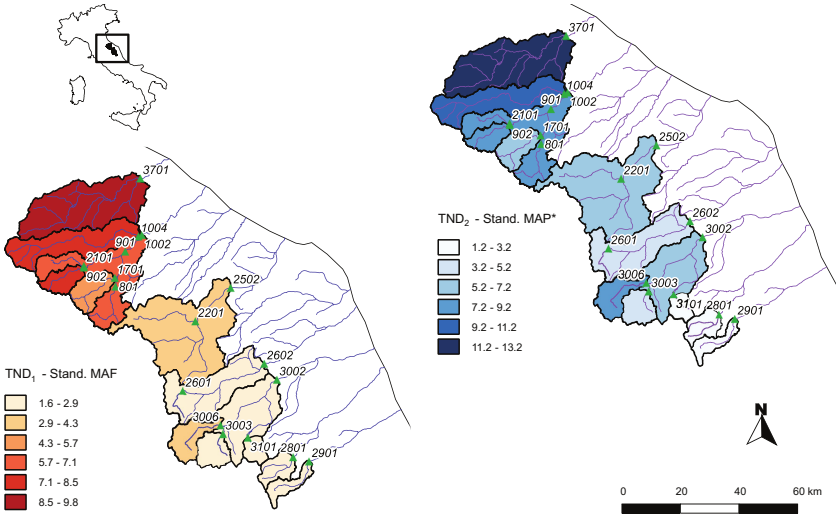

Fig. 3. Empirical TND_1 and TND_2 values for the study catchments.

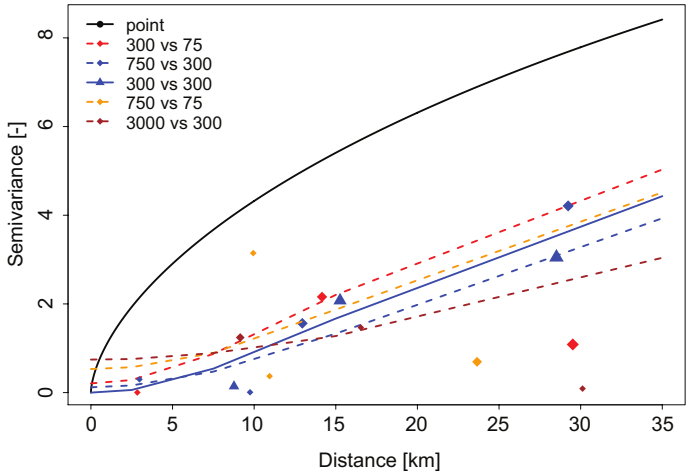

Fig. 4. ~~Empirical~~ Sample variogram (points) and ~~theoretical semivariograms from TND_1 values~~regularized variograms as function of distance and area. ~~Black~~ The black solid line ~~shows~~ represents the fitted point variogram~~. Colour markers and~~ , the blue line represents regularized variogram of equally sized catchments ($\sim 300\ \mathrm{km}^2$), dotted lines show ~~empirical and fitted variograms (empirical variograms are computed by binning catchment pairs for different~~ the effect of combinations of ~~catchment areas, e~~different catchments sizes in square kilometers (see also Fig. 4 in Skøien et al.~~g. $\simeq 300\,\mathrm{km}^2$ vs $\simeq 75\,\mathrm{km}^2$~~, 2014).

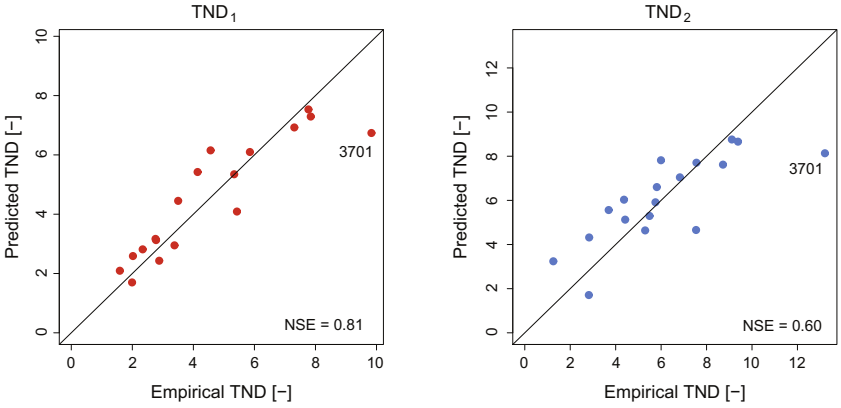

Fig. 5. Top-kriging predictions of TND_1 and TND_2 values in cross-validation, predictions for site 3701 are highlighted.

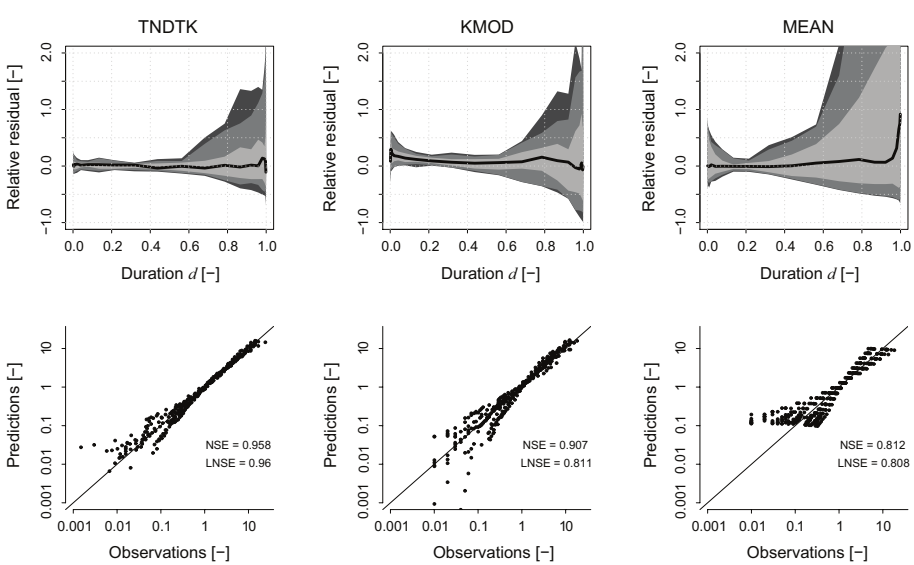

Fig. 6. Cross-validation of regional models: MEAN (right), KMOD (center), TNDTK (proposed approach, left); error-duration bands reporting the profile of the median relative error (thick black line) and the bands containing 50 %, 80 % and 90 % of the relative errors (grey nested bands) as a function of duration (top); empirical vs. predicted standardised streamflows (bottom).

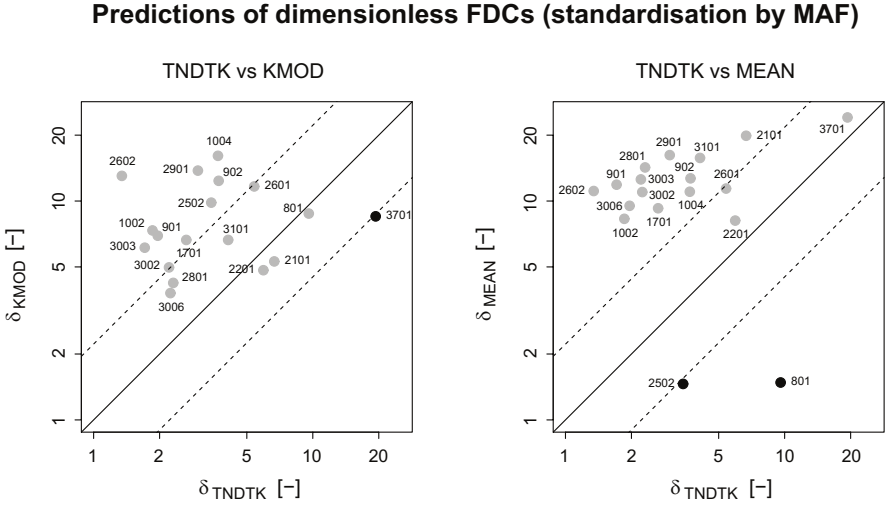

Fig. 7. Comparison between TNDTK, MEAN and KMOD models in terms of distances between empirical and predicted FDCs, δ_{mod} (where mod stands for TNDTK, MEAN or KMOD); values of δ_{TNDTK} are reported against values of δ_{KMOD} (left) or δ_{MEAN} (right) for each study basin; the solid line represents the ratio 1:1 between the errors, while in the area outside the dashed lines delimit the errors for the TNDTK model are twice as large as the MEAN or KMOD ones, or vice versa. Points above the solid line represent curves better estimated by TNDTK; points above the top dashed line represent curves that are much better estimated by TNDTK (see also Ganora et al.,2009, Fig. 8); sites 3701 and 801 are highlighted.

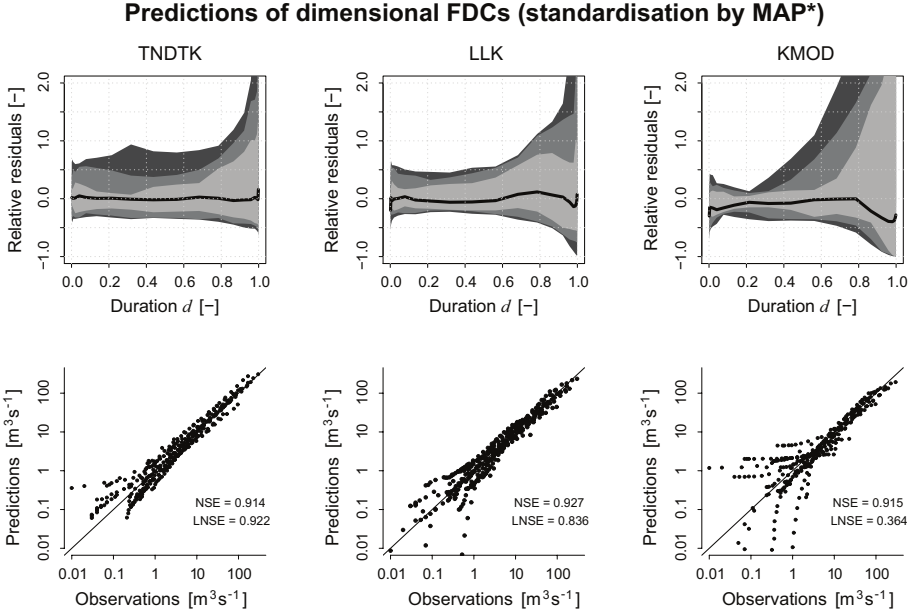

Fig. 8. Cross-validation of regional models: KMOD (right), LLK (center), TNDTK (proposed approach, left); error-duration bands reporting the profile of the median relative error (thick black line) and the bands containing 50 %, 80 % and 90 % of the relative errors (grey nested bands) as a function of duration (top); empirical vs. predicted dimensional streamflows (bottom).

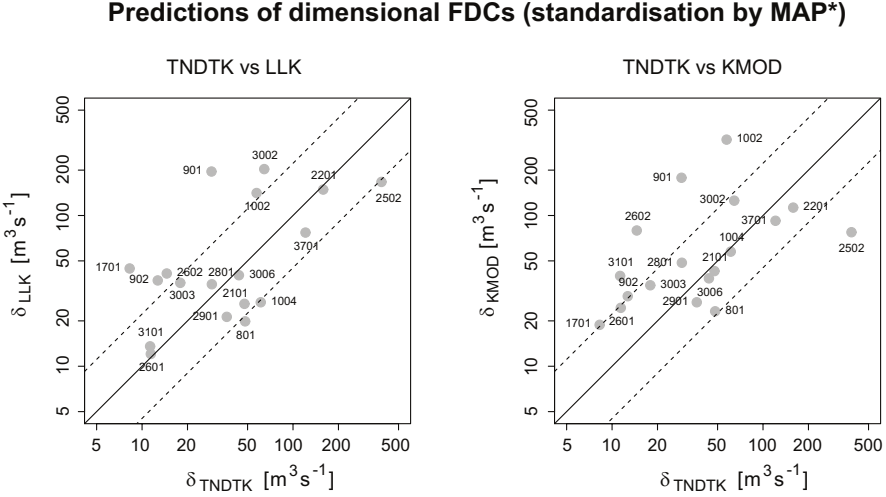

Fig. 9. Comparison between TNDTK, KMOD and LLK models in terms of distances between empirical and predicted dimensional FDCs, δ_{mod} (where mod stands for TNDTK, KMOD or LLK); values of δ_{TNDTK} are reported against values of δ_{LLK} (left) or δ_{KMOD} (right) for each study basin; the solid line represents the ratio $1:1$ between the errors, while in the areas outside the dashed lines delimit the areas where errors for the TNDTK model are twice as large as the LLK or KMOD ones, or vice versa. Points above the solid line represent curves that are better estimated by TNDTK; points above the top dashed line represent curves much better estimated by TNDTK (see also Ganora et al.,2009, Fig. 8).

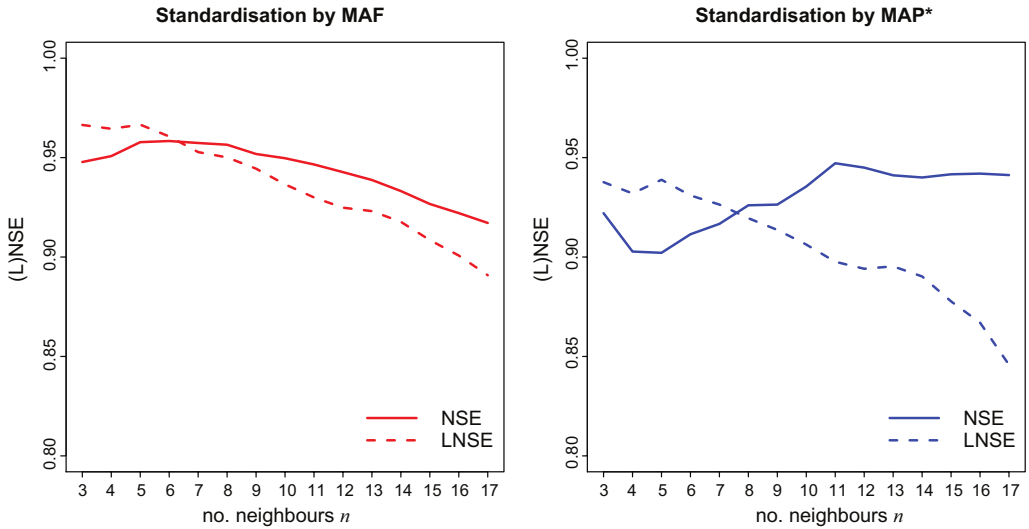

Fig. 10. Nash & Sutcliffe Efficiency for natural (NSE, filled lines) and log-transformed (LNSE, solid lines) streamflows plotted against the number n of neighbouring stations used for the interpolation. Left panel shows the predictions results for dimensionless FDCs (i.e. MAF standardisation), while the right panel reports the results for dimensional FDCs (i.e. MAP*).

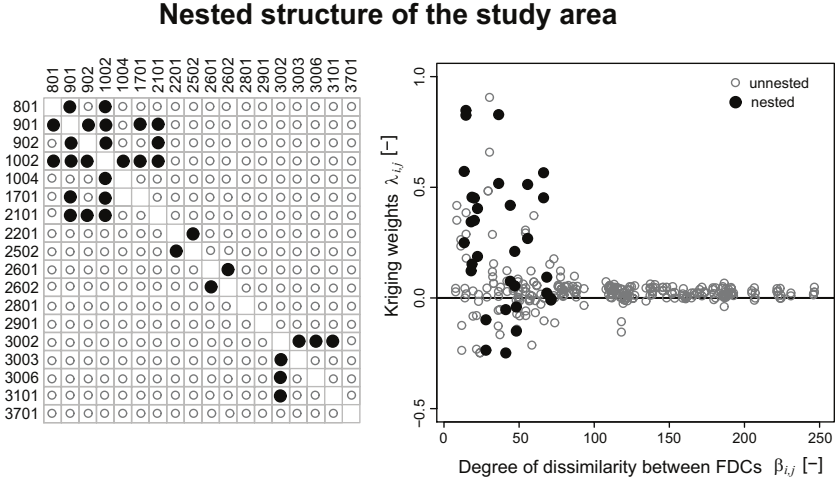

Fig. 11. Nested structure of the study area: (left) black dots identify nested pairs (i.e. basin-subbasin relationships); (right) Top-kriging weights $\lambda_{i,j}$ obtained for predicting TND_1 vs. the corresponding degree of dissimilarity between empirical FDCs for sites i and j, $\beta_{i,j}$, nested pairs are highlighted.

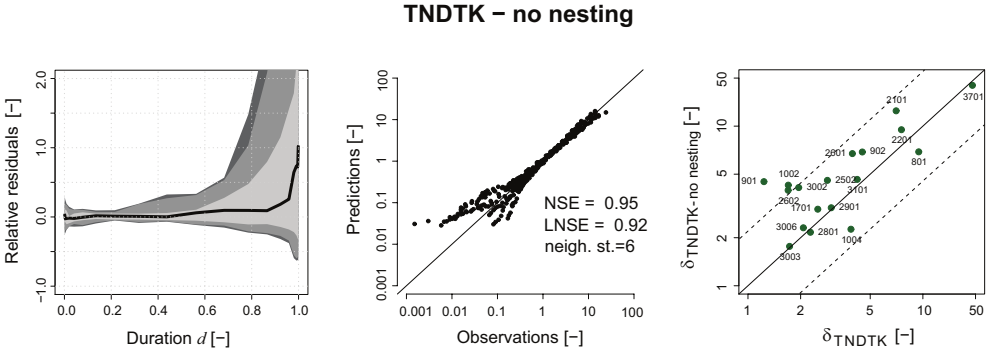

Fig. 12. Results of Leave Nested Out Cross-Validation (LNOCV): error-duration bands reporting the profile of the median relative error (thick black line) and the bands containing 50 %, 80 % and 90 % of the relative errors (grey nested bands) as a function of duration (left); empirical vs. predicted standardised streamflows (center); comparison of overall errors between empirical and predicted dimensionless FDCs, values of δ_{TNDTK} (Sec. 5.1.1) are reported against values of $\delta_{\text{TNDTK-no nesting}}$ (right).