An approach to propagate streamflow statistics along the river network

An approach to propagate streamflow statistics along the river network

D. Ganora, F. Laio and P. Claps

Department of Environment, Land and Infrastructure Engineering, Politecnico di Torino, I-10129 Torino, Italy

Abstract Streamflow at ungauged sites is often predicted by means of regional statistical procedures. The standard regional approaches do not preserve the information related to the hierarchy among gauged stations deriving from their location along the river network. However, this information is important when estimating runoff at a site located immediately upstream or downstream of a gauging station. We propose here a novel approach, referred to as the Along-Stream Estimation (ASE) method, to improve runoff estimation at ungauged sites. The ASE approach starts from the regional estimate at an ungauged (target) site, and corrects it based on regional and sample estimates of the same variable at a donor site, where sample data are available. A criterion to define the domain of application around each donor site of the ASE approach is proposed, and the uncertainty inherent in the estimates obtained is evaluated. This allows one to compare the variance of the along-stream estimates to that of other models that eventually become available for application (e.g. regional models), and thus to choose the most accurate method (or to combine different estimates). The ASE model was applied in the northwest of Italy in connection with an existing regional model for flood frequency analysis. The analysed variables are the first L-moments of the annual discharge maxima. The application demonstrates that the ASE approach can be used effectively to improve the regional estimates for the L-moment of order one (the index flood), particularly when the area ratio of a pair of donor–target basins is less than or equal to ten. However, in this case study, the method does not provide significant improvements to the estimation of higher-order L-moments.

Approche de la propagation des statistiques de débit le long d'un réseau hydrographique

Résumé Les débits des rivières dans les sections non-jaugées sont souvent estimés par des procédures statistiques régionales. Dans les méthodes régionales classiques aucune information relative à la hiérarchie géographique des stations placées le long du réseau hydrographique n'est retenue. Cette information est pourtant importante lorsqu'on estime des débits pour un site situé immédiatement en amont ou en aval d'une station de jaugeage. Nous proposons ici une nouvelle approche, appelée Estimation au fil de l'eau (EFE), afin d'améliorer l'estimation des débits aux sites non jaugés. L'approche EFE commence par l'estimation régionale en un site non jaugé (site cible), qui est ensuite corrigée à partir des estimations régionales et de l'échantillon de la même variable en un site donneur où les observations sont disponibles. Un critère particulier a été proposé pour définir le domaine d'application de l'approche EFE autour de chaque site donneur, ainsi que pour évaluer l'incertitude des estimations obtenues. Ceci permet de comparer la variance des estimations de l'EFE à celle des autres modèles statistiques éventuellement applicables (par exemple, les modèles régionaux classiques), et donc de choisir la méthode la plus précise (ou de combiner différentes estimations). Le modèle EFE a été appliqué dans le Nord-Ouest de l'Italie, dans le cadre d'une méthode existante d'analyse régionale de probabilité des crues. Les variables estimées aux sites non jaugés sont les premiers L-moments de la crue annuelle. L'application démontre que l'approche EFE peut être utilisée efficacement afin d'améliorer les estimations régionales du L-moment d'ordre un (l'indice de crue). Ceci est vrai en particulier lorsque le rapport des surfaces des bassins d'une paire donneur-cible est inférieur ou égal à dix. Dans notre étude de cas la méthode ne démontre pas d'amélioration importante de l'estimation des L-moments d'ordre supérieur.

1 INTRODUCTION

Prediction of streamflow statistics in ungauged basins is often performed through the use of regional models (e.g. Grimaldi *et al.* 2011); a procedure able to exploit local information and to improve regional estimates would thus be useful for many purposes.

Several types of regional models have been proposed in the literature based on the underlying idea to "substitute time for space" (US National Research Council 1988), i.e. to compensate for the lack of data record at a certain location by transferring the information from other gauged sites. These models differ in terms of the regionalized variable and the mathematical framework used for the information transfer, while their common focus is to consider the transfer of hydrological information moving to the so-called descriptor space. The dimensions of the descriptor space are the catchment characteristics (usually topographic, morphological, pedological or climatic) that can be computed for each basin without resorting to hydrologic data. Then, suitable relationships are built to relate some of these characteristics to the desired hydrological variable, thus providing a tool for estimating the variable in ungauged basins.

Differing from regional approaches, the basic concept underpinning the model developed in this work is the transfer of hydrological information to an ungauged site located upstream or downstream of existing gauging stations. This "propagation of information" involves a supporting variable calculated at a gauged (or donor) basin on the basis of sample data that are used to propagate the information towards the ungauged (target) site. The target and the donor site are directly connected by the drainage network, i.e. the two drainage basins are nested.

Given this perspective, this information transfer can be supposed to be helpful only if the two sites are close enough. The estimation of the uncertainty of the propagated prediction is thus a key element, since it allows one to evaluate whether the propagated prediction is better than the regional model prediction. This model is named Along-Stream Estimation (ASE) in the following, to underline that it is based on the river network structure. Any discharge-related variable can in principle be propagated with the ASE approach. The procedure is applied here to the first L-moments (e.g. Hosking and Wallis 1997) computed on the record of annual streamflow maxima. These statistics can be profitably used to estimate flood frequency curves (Laio *et al.* 2011).

The issue of prediction or interpolation of hydrological variables along the river network is not frequently discussed in the literature, although some notable exceptions are found. Kjeldsen and Jones (2007) adopted a procedure which, analogous to our model, implements the idea of locally correcting the regional estimates on the basis of proximal sources of information. However, this procedure does not account for the river network structure. We will return to the similarities and differences with the work of Kjeldsen and Jones (2007) later in this paper (Section 4).

Gottschalk (1993a, 1993b) approached a similar problem considering the network structure and introduced the issue of correlation and covariance of runoff, adapting the theory of stochastic processes to the hierarchical structure of nested catchments. This approach was later extended by Gottschalk *et al.* (2006, 2011), and similar concepts are used by Skoien *et al*. (2006) in the development of a kriging procedure that accounts for the river structure, named topological kriging or top-kriging. These approaches, differing from the ASE model, do not aim to correct regional estimates already available for the ungauged sites, but provide independent estimators of design floods.

2 ALONG-STREAM INFORMATION PROPAGATION METHOD

2.1 Method and assumptions

In the ASE approach, a generic hydrological variable (e.g. the mean of annual streamflow maxima) is represented as S when it is computed on the empirical sample, while the same variable computed through the propagation of information is denoted by P. Moreover, to allow generalizations, the gauged or donor site is denoted by the subscript d, and the ungauged or target site by the subscript t.

The approach is based on the following hypotheses:

- Proximity: the target site is located on the same stream path as the donor station, upstream or downstream, i.e. the two basins d and t are nested.
- Transferability: the variable S_d, computed at the donor site, is used as the support of propagation in the information transfer scheme, $P_t = f(S_d,\theta)$, where P_t is the propagated variable at the target site, θ is an additional (optional) set of parameters and f is a function to be defined.
- Congruence: when the distance between the donor and the ungauged catchments becomes null, the

two basins coincide and the along-stream estimate at the ungauged site matches the at-site estimate at the gauged basin, i.e. $P_t \rightarrow S_d$ for $t \rightarrow d$. The distance is intended with a general meaning, and does not necessarily represent the geographical distance, or the length of the drainage path between the two points.

To set the validity domain is very important for assessing the reliability of the ASE method. Here the point will be treated in an intuitive way: the idea is to consider the ASE model applicable only where the uncertainty, i.e. the standard deviation, σ_{Pt}, of the propagated prediction can be suitably estimated. The σ_{Pt} can be statistically evaluated by considering the residuals obtained during the calibration phase of the method: the more efficient the function for the information transfer, the smaller are the residuals, and the larger is the domain of validity.

The prediction, P_t, is obtained through the function f, and its standard deviation can be obtained as the combination of a model error, due to the approximate form of f, and a sample error, owing to the use of the variable S_d, which is computed from the data. Considering these sources of uncertainty in detail is out the scope of this paper; however, the variance of P_t can be quantified in a simplified way: given the particular function f for the information transfer, together with its corresponding domain of validity, the variance of the along-stream prediction is assumed to increase moving away from the donor site, but still within the validity domain. Outside this domain, the along-stream prediction is deemed unreliable and it is therefore no longer necessary to compute its variance. A sketch representing this aspect is shown in Fig. 1(c) and details are reported in Section 2.3.

The ASE approach is meant to be applied to an area for which a regional flood frequency model has

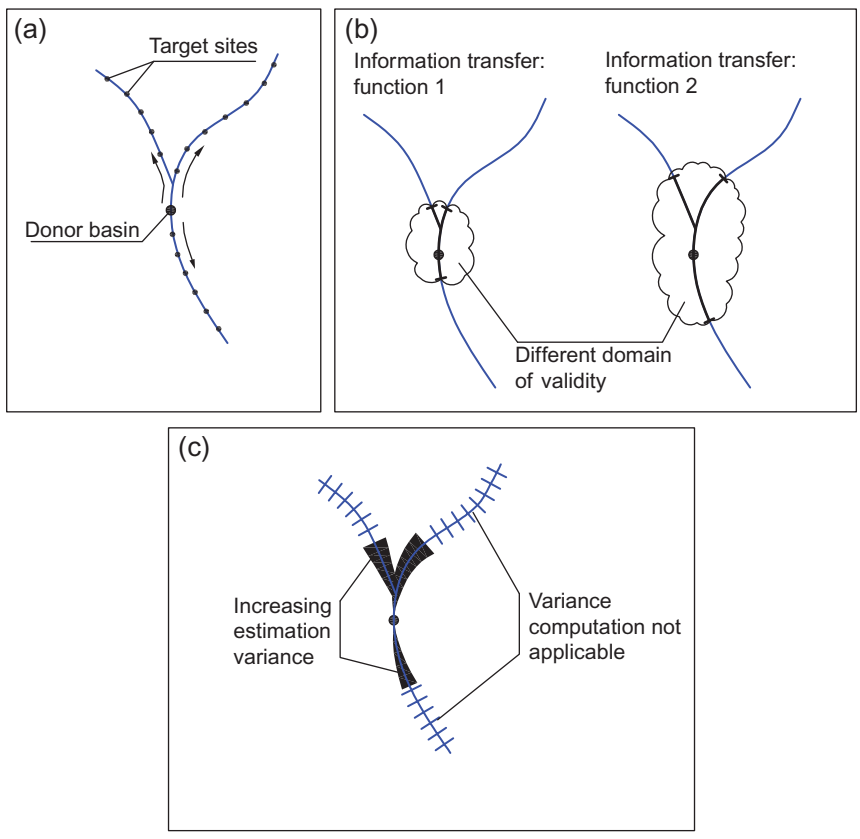


Fig. 1 Sketch of the along-stream propagation of information: (a) a hydrological variable calculated at the donor (gauged) site is used to predict the value of the same variable at the target locations located upstream or downstream. (b) Different functions can be adopted to achieve this aim; however, each function has a particular domain of validity around the donor station. (c) The variance of the new predictions is assumed to increase moving away from the donor station, within the domain of validity. This is no longer applicable out of the validity domain.

already been developed. The regional model is used: (a) as a reference model for results comparison; (b) as a substitute for P_t where the ASE approach is not applicable; and (c) as a source of information for the definition of the additional set of parameters θ. For instance, the regional model used here (presented in detail by Laio *et al.* 2011) allows one to evaluate the flood frequency curve in ungauged basins through the estimation of three L-moments on the basis of a set of three regional relationships: the index flood (Q_{ind}, average of annual maxima), the coefficients of L-variation (L_{CV}) and L-skewness (L_{CA}).

2.2 Propagation of information

The first step to implement the along-stream estimation procedure is to define a suitable function f to compute the variable P at the target site t, given the value S_d at the donor site. Here we adopt an equation proposed in the *Flood Estimation Handbook* (Institute of Hydrology 1999) and re-analysed by Kjeldsen and Jones (2007). The function f reads:

$$f(S_d, \theta) = \frac{R_t}{R_d} S_d \tag{1}$$

where R refers to the estimates obtained from the regional procedure and S_d is the at-site sample value of the variable at the donor site. Equation (1) can be interpreted as follows: the regional estimate R_t in t is corrected by a factor equal to the relative error that the regional model produces in d (i.e. S_d/R_d). Note that here all the symbols P, R and S represent the same hydrological variable of interest (e.g. the index flood, L_{CV}, or L_{CA}).

The propagated estimate of P_t can then be written as:

$$\begin{cases} P_t = \dfrac{R_t}{R_d} S_d & \text{if } D \le D_{\text{lim}} \\ P_t = R_t & \text{if } D > D_{\text{lim}} \end{cases} \tag{2}$$

where D is the generalized distance between t and d and D_{lim} is the threshold distance beyond which the propagation is no longer effective. For $D \to 0$ it is straightforward to verify that $P_t \to S_d$. In this context, since we already have an alternative model (the regional model) available for the prediction of the variable at the ungauged site, it seems appropriate to use the pure regional estimates in the cases where $D > D_{\text{lim}}$.

2.3 Model reliability: operational estimate and prediction uncertainty

The framework introduced in Section 2.1, which highlights the idea that the model is applicable only in a limited neighbourhood, is common also to other approaches for local correction of regional estimates. However, in our methodology, we explicitly evaluate the effectiveness of such correction. From a practical point of view, this introduces a further rule with respect to the propagated estimate of equation (2) that is formalized by defining the operational or along-stream estimate ASE_t as:

$$\begin{cases} \text{ASE}_t = P_t & \text{if } D \le D_{\text{lim}} \text{ and } \sigma_{P_t} \le \sigma_{R_t} \\ \text{ASE}_t = R_t & \text{otherwise} \end{cases} \tag{3}$$

where σ_{R_t} is the standard deviation of the regional prediction, and σ_{P_t} is the standard deviation (to be evaluated) of the propagated variable. This means that, even for $D \le D_{\text{lim}}$, the propagated prediction P_t is accepted only if σ_{P_t} is not greater than the corresponding regional uncertainty σ_{R_t}. The standard deviation of the regional prediction should be available from the regional model used (see e.g. Laio *et al.* 2011, for our case study).

The basics of the ASE method can then be summarized in three steps: (a) choice of a suitable framework for the information transfer, (b) definition of the threshold distance, D_{lim}, and (c) evaluation of the uncertainty of the propagated estimate. In Section 2.2, a practical formula (equation (2)) is proposed without providing a quantitative assessment of D_{lim}. In this section, we investigate the suitability of a simplified approach for D_{lim} quantification (point (b)), together with an overall evaluation of the performance of the along-stream estimation approach (point (c)). These two steps are operated jointly through an iterative procedure to provide an optimal estimator of D_{lim}. This iterative framework is general, and can also be applied to propagate equations that are different from equation (2).

To implement the ASE framework, it is necessary to define a suitable relationship that represents the uncertainty of P_t. As mentioned earlier, an analytical equation for σ_{P_t} could be derived on the basis of equation (1), although actually the effect of the model error on σ_{P_t} is not easy to define. Consequently, in our approach, we resort to a simple model of the P_t uncertainty:

$$\text{CV}_{P_t} = (1 + \alpha D)\text{CV}_{S_d} \tag{4}$$

where CV is the coefficient of variation of the propagated variable, i.e. the ratio between the standard deviation and the mean of the variable. This model for predicting CV_{P_t} can be interpreted as follows: the coefficient of variation of P_t equals the coefficient of variation of the at-site estimate in the gauged site (which includes the sample uncertainty) augmented proportionally to a factor α that accounts for both the non-correctness of the ASE transfer function (model error) and for the variance of the other variables involved in equation (2) (in this case the regional values R). This can also be thought of as a first-order approximation of a more complex function for variance propagation.

Considering the definition of P_t given in equation (2), and the definition of CV in equation (4), we obtain for $D \leq D_{\text{lim}}$:

$$\sigma_{P_t} \frac{R_d}{R_t S_d} = (1 + \alpha D) \frac{\sigma_{S_d}}{S_d} \tag{5}$$

and thus:

$$\sigma_{P_t} = (1 + \alpha D) \sigma_{S_d} \frac{R_t}{R_d} \tag{6}$$

For $D \to 0$, it is straightforward to verify that $\sigma_{P_t} \to \sigma_{S_d}$, confirming the congruence hypothesis.

The evaluation of the uncertainty of the propagated estimate using equation (6) first requires estimation of the parameter α. As a first attempt, we calibrated α on the basis of the available data set, rearranged to account for each donor–target correspondence (details in Section 3) and considering only the basin pairs within the threshold distance. Given that, for each pair of basins, the residual between P_t and its corresponding at-site value S_t is:

$$\delta_t = P_t - S_t \tag{7}$$

P_t and S_t are assumed to be independent random variables neglecting the covariance between S_d and S_t. This hypothesis allows one to keep the framework simple and is justified by the fact that the covariance of flood statistics rapidly declines in orographically complex areas and cannot be robustly estimated in the study area (Laio *et al*. 2011). Using equation (6) one obtains:

$$\begin{aligned} \sigma_\delta^2 &= \sigma_{P_t}^2 + \sigma_{S_t}^2 \\ &= (1 + \alpha D)^2 \sigma_{S_d}^2 \left(\frac{R_t}{R_d} \right)^2 + \sigma_{S_t}^2 \end{aligned} \tag{8}$$

The coefficient α can thus be estimated by means of a maximum-likelihood approach: the residuals δ_t are supposed to be normally distributed with zero mean and variance changing site-by-site accordingly to equation (8). The likelihood function is obtained as the product of each (normal) marginal probability distribution.

Note that, even if equation (6) relates a standard deviation to a distance like a variogram, it is conceptually different from a variogram because the calibration is performed using the residuals δ_t, which does not require the availability of simultaneous observations. This is particularly important in the context of limited data availability, where the variograms cannot be robustly estimated.

The value of D_{lim} and α (which are not known *a priori*) are optimized by means of a trial-and-error procedure:

(a) a tentative value of D_{lim} is selected;
(b) the propagated estimate P_t is computed as in equation (2);
(c) the residuals δ_t are computed and the parameter α is evaluated in the max-likelihood framework, only for basin pairs within D_{lim};
(d) based on α, the variance of the P_t prediction is computed with equation (6) and it is compared against the variance of the regional prediction at the same location;
(e) the operational estimate ASE_t is obtained by following the rules in equation (3);
(f) an error index (see equation (11) in the following) is computed both for the ASE model and the regional model, and the two error indexes are compared;
(g) the procedure is repeated, changing the tentative value of D_{lim}; and
(h) the D_{lim} value that minimizes the overall error of ASE_t is assumed as the distance threshold.

This procedure considers a unique D_{lim} value which is valid for the whole case study. A "global" D_{lim} value is necessary, from the computational point of view, to perform the estimation of α. The search for D_{lim} is based on a trial-and-error procedure that roughly tells us the maximum distance wherein sensible improvements are obtained. Nevertheless, precise

estimation of D_{lim} would be superfluous, because, even when $D < D_{\text{lim}}$, the propagated estimate is further compared to the regional one to choose the most appropriate estimate (see equation (3)).

3 CASE STUDY

3.1 Organization of nested basins

The regional models employed in this work are in the form of multiple regression models calibrated over a data set of 70 basins located in northwest Italy; geomorphologic and climatic indexes available for any basin are used as explanatory variables (Laio *et al.* 2011). The basins belong mainly to mountainous areas, have areas ranging between 22 and 3320 km^2 and mean elevation from 471 to 2719 m a.s.l. To reduce any effect of upstream lakes and/or reservoirs, basins whose catchment area is more than 10% covered by lakes are discarded. The investigated region presents basins subject to various climate regimes, from purely nivo-glacial to almost temperate-Mediterranean. Further details can be found in Claps and Laio (2008). Each prediction model takes the form:

$$\hat{Y} = \boldsymbol{x}^T \hat{\boldsymbol{\beta}} \tag{9}$$

where $\hat{Y}$ is the regionalized variable (L-moment), $\boldsymbol{x}$ is the vector containing the descriptors for the considered basin, with one as the first element, and $\hat{\boldsymbol{\beta}}$ is the vector of regression coefficients, obtained following a generalized least-squares procedure modified after Stedinger and Tasker (1985). Laio *et al.* (2011) found that, for the estimation of the index flood, it is more appropriate to apply equation (9) to the log-transformed data; consequently, a back-transformation is required to obtain Q_{ind} from $\hat{Y}$. The flood frequency curve is then reconstructed considering the regionalized L-moments. The regional model of Laio *et al.* (2011) also allows one to evaluate the standard deviation of each L-moment predicted at an ungauged site.

The suitability of the ASE approach is evaluated considering as a case study the same set of 70 basins used by Laio *et al.* (2011); however, here the data are organized in a different way, it being more appropriate to work in terms of pairs of basins $\{t,d\}$, rather than with a single catchment at a time. Figure 2 shows a schematic representation of the hierarchical dependence of nested catchments, the connections being represented by a line. Note that there are several multi-connected basins, as well as basins with no connections. All the connected (nested) catchments have been considered as possible pairs of donor–target sites (e.g. in Fig. 2, Basin 1 is nested to Basin 15 although Basin 13 occurs in between the two).

The connections are considered *in both directions*, e.g. if Basin 9 is upstream of Basin 10, we first consider Basin 9 as the donor site and Basin 10 as the target (ungauged) site; then the procedure is repeated using Basin 10 as the donor site and Basin 9 as the target (ungauged) site. Considering all of the possible connections of two stations along the same drainage path (nested basins), there is a total of 142 connections. Every pair is characterized by a generalized distance D between them.

The distance D between two catchments can be defined in different ways, but it is preferable to avoid both the geographical distance and the length of the drainage path linking the two closing sections. For instance, such definitions would not represent correctly the abrupt change in basin characteristics that is expected between two points located just upstream or downstream of a tributary. We propose a definition of the distance based on the ratio of basin areas A:

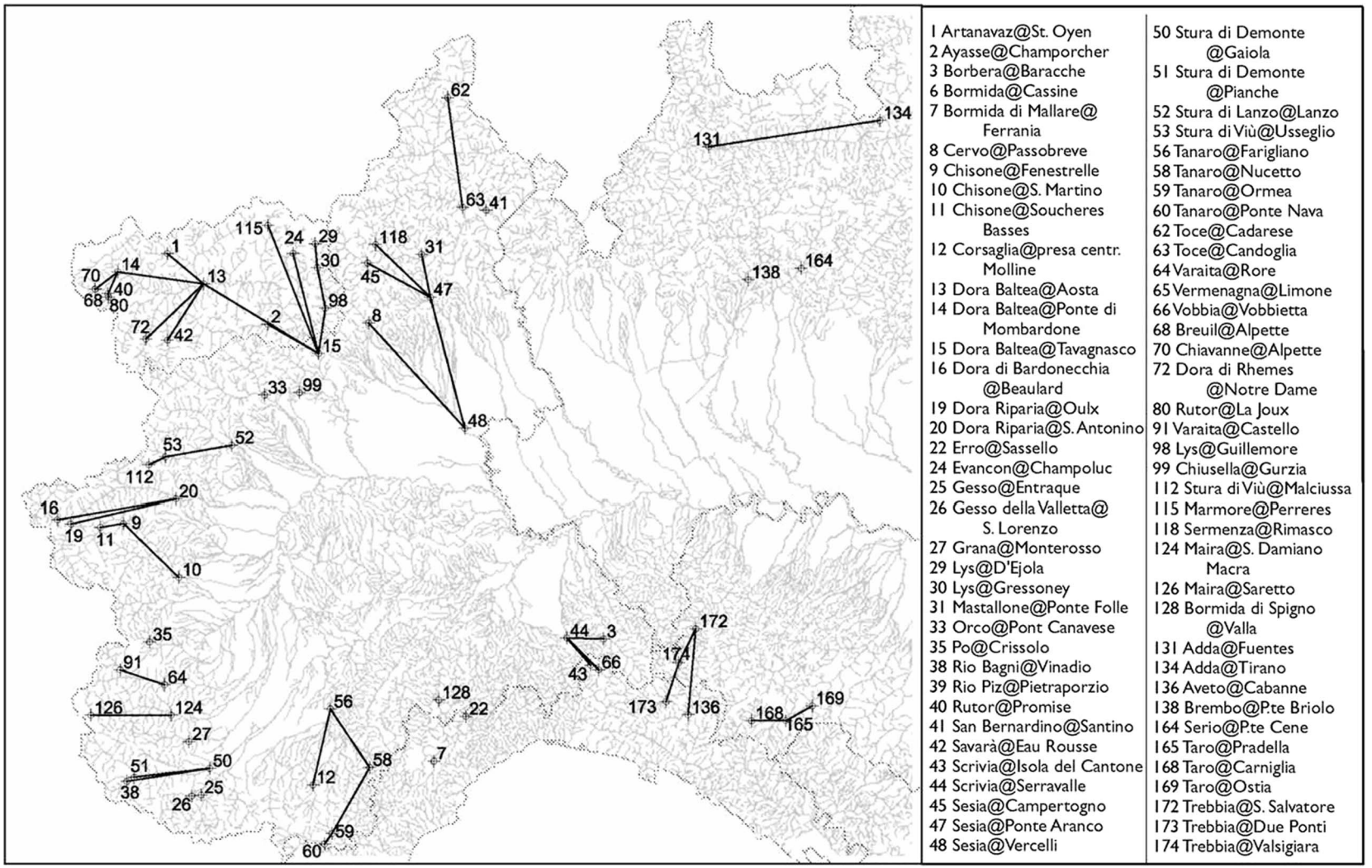
$$D = \log(A_{\max}/A_{\min}) \tag{10}$$

with $A_{\max} = \max[A_t, A_d]$ and $A_{\min} = \min[A_t, A_d]$. Under the proximity hypothesis (but not in general), two basins with the same area have null distance (they are the same basin). Consequently, their estimates must coincide (congruence hypothesis). Other variables may be included in the representation of the generalized distance; for example, the mean basin elevation can be useful when the data set is composed of basins from both mountainous and plain areas.

3.2 Application of the iterative procedure

The trial-and-error procedure described in Section 2.3 was applied to the index flood, the L_{CV} and the L_{CA} statistics, using equation (10) as the distance measure between basins. The variable $\log(Q_{\text{ind}})$, i.e. the log-transformed index flood, was also considered, using a slightly modified version of the ASE approach in order to compare our model with that proposed by Kjeldsen and Jones (2007). Details of this comparison are reported in Section 4.

The main results of the application of the ASE method with a number of tentative threshold distances are summarized for the index flood in Fig. 3, where



1 Artanavaz@St. Oyen
2 Ayasse@Champorcher
3 Borbera@Baracche
6 Bormida@Cassine
7 Bormida di Mallare@ Ferrania
8 Cervo@Passobreve
9 Chisone@Fenestrelle
10 Chisone@S. Martino
11 Chisone@Soucheres Basses
12 Corsaglia@presa centr. Molline
13 Dora Baltea@Aosta
14 Dora Baltea@Ponte di Mombardone
15 Dora Baltea@Tavagnasco
16 Dora di Bardonecchia @Beaulard
19 Dora Riparia@Oulx
20 Dora Riparia@S. Antonino
22 Erro@Sassello
24 Evancon@Champoluc
25 Gesso@Entraque
26 Gesso della Valletta@ S. Lorenzo
27 Grana@Monterosso
29 Lys@D'Ejola
30 Lys@Gressoney
31 Mastallone@Ponte Folle
33 Orco@Pont Canavese
35 Po@Crissolo
38 Rio Bagni@Vinadio
39 Rio Piz@Pietraporzio
40 Rutor@Promise
41 San Bernardino@Santino
42 Savarà@Eau Rousse
43 Scrivia@Isola del Cantone
44 Scrivia@Serravalle
45 Sesia@Campertogno
47 Sesia@Ponte Aranco
48 Sesia@Vercelli
50 Stura di Demonte @Gaiola
51 Stura di Demonte @Pianche
52 Stura di Lanzo@Lanzo
53 Stura di Viù@Usseglio
56 Tanaro@Farigliano
58 Tanaro@Nucetto
59 Tanaro@Ormea
60 Tanaro@Ponte Nava
62 Toce@Cadarese
63 Toce@Candoglia
64 Varaita@Rore
65 Vermenagna@Limone
66 Vobbia@Vobbietta
68 Breuil@Alpette
70 Chiavanne@Alpette
72 Dora di Rhemes @Notre Dame
80 Rutor@La Joux
91 Varaita@Castello
98 Lys@Guillemore
99 Chiusella@Gurzia
112 Stura di Viù@Malciussa
115 Marmore@Perreres
118 Sermenza@Rimasco
124 Maira@S. Damiano Macra
126 Maira@Saretto
128 Bormida di Spigno @Valla
131 Adda@Fuentes
134 Adda@Tirano
136 Aveto@Cabanne
138 Brembo@P.te Briolo
164 Serio@P.te Cene
165 Taro@Pradella
168 Taro@Carniglia
169 Taro@Ostia
172 Trebbia@S. Salvatore
173 Trebbia@Due Ponti
174 Trebbia@Valsigiara

Fig. 2 Gauging stations lying on the same drainage path, either upstream or downstream, that are directly connected (nested basins) are schematically linked with a line. Some catchments have multiple connections, others are isolated.

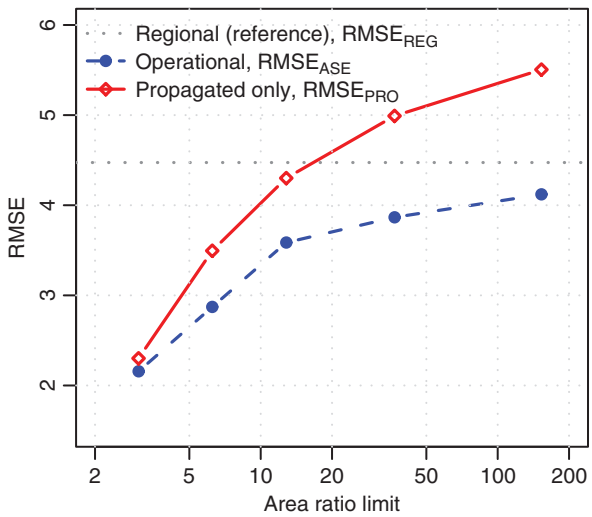


Fig. 3 Average error (RMSE of dimensionless errors $\varepsilon_{\{t,d\}}$) within the domain of validity considering only the propagated estimate (——) and the operative ASE prediction (- - - -). The global average error of the regional model is indicated, for reference, by a dotted line.

the root mean squared errors (RMSE) of the normalized prediction errors are plotted as a function of the area ratio of the donor and target basins. The normalized errors are defined as:

$$\varepsilon_{\{t,d\}} = \frac{(\text{prediction})_{\{t,d\}} - S_t}{\sigma_{S_t}} \tag{11}$$

where "prediction" indicates the approach used to make the estimation; the residuals are normalized by σ_{S_t} to account for the sample uncertainty at the target site, which can be relevant if the donor has a short record.

In the first instance, Fig. 3 allows one to compare the behaviour, in terms of RMSE, of the operational ASE estimator of equation (3) (in the following referred to as RMSE$_{ASE}$), against the simple propagation of information of equation (2) (RMSE$_{PRO}$). Both approaches were applied over all possible pairs $\{t,d\}$, but only within the distance limit. The last point of the trial-and-error procedure is relative to a distance limit of 5.03 (equivalent to an area ratio of about 150) that includes all the available basin pairs, i.e. it is equivalent to an unbounded domain of validity. In Fig. 3, the global RMSE computed considering only the regional predictions over the whole data set, RMSE$_{REG}$, is also reported for comparison.

Some important results can be deduced from the RMSE$_{PRO}$ curve: it presents a clear increasing trend, with increasing threshold distance; and RMSE$_{REG}$ is equalled for an area ratio between 10 and 20. This indicates that the use of a simple propagation of information as in equation (2) is effective only for relatively short distances.

More important, Fig. 3 shows the effectiveness of the ASE method relative to the simple propagation approach; in fact, the RMSE$_{ASE}$ is always lower than the RMSE$_{PRO}$, meaning that the selection criterion in equation (3), based on the standard deviation of the propagated and regional estimates, works properly; in other words, this is confirmation that, on average, the operational model is able to correctly select the best approach (regional or propagated). As expected, for large area ratios, the ASE performances approach those of the regional model because σ_{P_t} increases and the regional model is selected most of the time in equation (3). Thus, the ASE model has better performances compared to the regional model alone, even when there is no distance limit.

These results highlight that the use of a restricted domain of validity improves the effectiveness of the propagation of information and, as a consequence, the whole ASE framework. However, a restricted domain of validity limits the applicability of the ASE method to only the closest target basins.

The optimal threshold distance can thus be seen as the best compromise between two opposite effects: on the one hand, the use of a small threshold distance D_{lim} leads to better estimation results, but the applicability of the ASE approach turns out to be limited to only a few basins. On the other hand, larger domains of validity increase the errors and decrease the effectiveness of the operational estimator.

The search for an optimal D_{lim} value has been performed iteratively for this case study, considering the calibration set of a basin as representative of the real application context. For instance, very good performances can be achieved with $D_{\text{lim}} = 0.81$ (equal to an area ratio of about 2.25), but only 11.3% of the considered basins would benefit in this case of the along-stream model. The remaining 88.7% of the basins would not be considered. Given this perspective, we selected the "optimal" distance as a balance between these two effects; this corresponds to extending the area of influence to basins that have an area of between 1/10 and 10 times the area of the donor basin, i.e. for pairs of basins whose areas differ by, at most, one order of magnitude.

The results reported in Fig. 3 show the global performances of the method. A more detailed investigation is represented in Fig. 4(a), in which each normalized error of the operational model is compared to that of the regional (reference) model. The

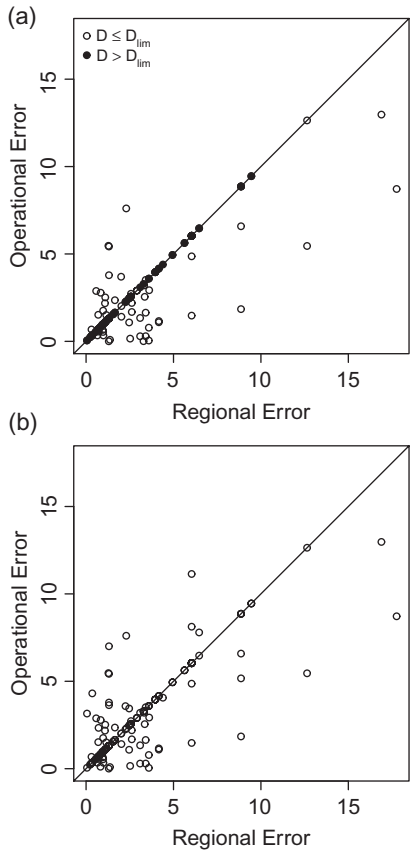


Fig. 4 Absolute errors (equation (11)) of regional estimates of the index flood compared with the errors produced by the ASE model. Open circles represent the errors obtained for basin pairs closer than the threshold distance; filled circles are relative to the more distant catchments. All the points below the solid line represent basins where the index flood estimates are improved by the use of the along-stream information transfer procedure: (a) relative to a threshold distance $D_{\text{lim}} = \log(10)$; (b) no limitation on distances.

points on the graph can be divided in four different classes:

- filled circles on the bisector represent basins outside the validity domain, where only the regional model is applicable;
- empty circles on the bisector are basins within D_{lim} for which the regional model has been selected as the operational model;
- empty circles below the bisector are basins within D_{lim} for which the propagated estimate has been selected, and the propagated estimates provide an improvement over the regional ones;
- empty-circles above the bisector are basins within D_{lim} for which the propagated estimate has been selected, but the propagated errors are greater than the regional ones.

Most of the points off the diagonal are in the lower part of the plot, which demonstrates that, when the propagated estimate is suitable for use, it provides better performances than the corresponding regional estimation. Only for a few basins is there a moderate increase in the operational error. These results are positive when compared to those of Fig. 4(b), in which no threshold distance was applied. Although the comprehensive operational error (RMSE_{ASE}) still suggests use of the ASE model, the dispersion of the points highlights the fact that the variance of the operational predictions is no longer appropriate to describe the reliability of the ASE model. This again confirms that, for basins beyond the threshold distance, the regional model is the most appropriate.

The same procedure was applied for the L_{CV} and L_{CA} estimations, but no conclusive results were reached. Figure 5 clearly shows, for L_{CV}, that the ASE model does not produce reliable results and, when applicable, produces a deterioration of the regional estimates. Similar results apply to L_{CA}, for which the method seems not applicable at all. These negative

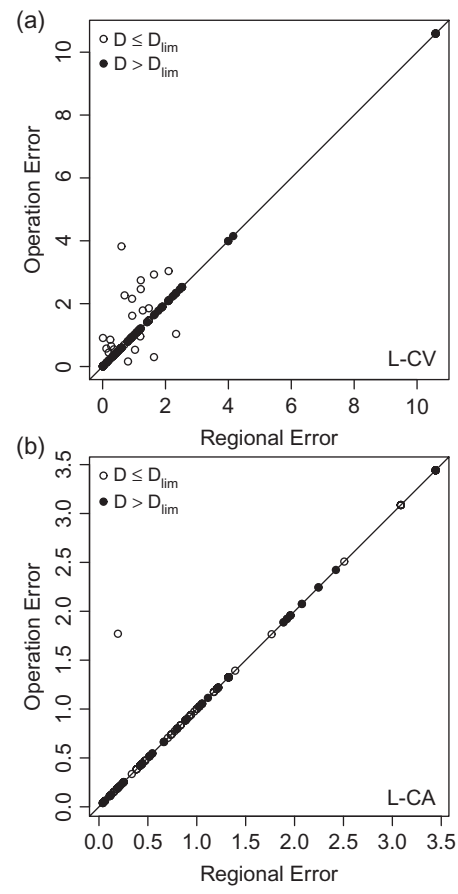


Fig. 5 ASE *vs* regional errors for (a) L_{CV} and (b) L_{CA} with optimal threshold distance $D_{\text{lim}} = \log(5)$.

results can be ascribed in part to the high uncertainty of the sample higher-order L-moments estimated on short data records. This uncertainty prevents correct estimation of the parameter α and of the bounds of the validity domain, thus deteriorating the quality of the results obtained with the ASE approach: the same effect influences the size of the domain of validity of the ASE approach, with D_{lim} decreasing with increasing order of the L-moment. In our case study, the domain of validity becomes so small that there are not enough pairs of basins included within the threshold distance that can be used for a robust calibration. Lack of data not only affects the sample uncertainty, but also makes it difficult to investigate the complex mechanisms of propagation of the second- and third-order L-moments. The available database cannot support a detailed analysis of such mechanisms, making the uncertainty related to the "model error" impossible to estimate, and hampering the applicability of the procedure.

4 MODEL COMPARISON

Kjeldsen and Jones (2007) developed a similar approach (hereafter the KJ approach) to locally improve the predictions coming from a regional model. This approach has been rediscussed (Kjeldsen and Jones 2009) and applied also in Kjeldsen and Jones (2010). Although the equation we use to transfer the information is basically the same as that of the KJ model, the two implementations are based on rather different ideas. In particular, Kjeldsen and Jones (2007) propose the model:

$$P_t = R_t \left(\frac{S_d}{R_d} \right)^{\alpha_{\text{KJ}}} \tag{12}$$

where α_{KJ} is an exponent dependent on the geographical distance of the centroids of the donor and target basins. The donor basin is always selected as the geographically-closest gauged basin. To evaluate the suitability of these approaches for application in the present case study, a comparison was carried out.

To evaluate α_{KJ}, the KJ model requires the estimation of the cross-correlation coefficient of the model errors. As a first approximation, and for practical purposes (see Kjeldsen and Jones 2007), it can be assumed that α_{KJ} depends on the distance from the donor site following the cross-correlation of annual maxima $r_{t,d}$. This approach applies to all the target sites, even if for large donor–target distances the correction is negligible, because α_{KJ} tends to zero. A special case of equation (12), reported by Kjeldsen and Jones (2007), considers $\alpha_{\text{KJ}} = 1$, provided the correction applies only for basins within a limit-distance (i.e. only for highly-correlated basin pairs). Beyond the limit-distance, defined on the basis of the correlation function, only the regional model is used.

In our case study, the regional model does not provide the cross-correlation function of the model errors, and the cross-correlation of annual maxima cannot be safely estimated over the considered area because the samples used are sparse in space and not completely overlapping in time. Moreover, the cross-correlation function of annual maxima is expected to decay very quickly, due to the high topographic and climatic heterogeneity in the case study area. To overcome this problem, an iterative procedure is adopted to calibrate the KJ model: the limit-distance is assumed to be known, with varying values from 1 to 200 km; the model is applied correcting only the within-limit pairs of target–donor basins; finally a comprehensive error index is computed. In this way, the most appropriate limit-distance is found to be 8 km, which is the distance that allows one to improve most of the estimates. This limit allows us also to roughly reconstruct the correlation function in the form of a negative exponential. Kjeldsen and Jones (2007) found the correlation function $r_{t,d} = \exp(-0.016 D_C)$ (D_C being the distance between basin centroids) valid for their case study, with the maximum distance for which the model applies corresponding to $r_{t,d} = 0.5$. Assuming this value is valid also in our case study, and considering the limit-distance of 8 km, the correlation function is re-evaluated as $r_{t,d} = \exp(-0.087 D_C)$, showing a faster decay than the Kjeldsen and Jones case study (which may be sensible, due to the larger meteorological variability in the study area compared to the UK). This result is necessary for applying the general version of the KJ model (equation (12)).

At this point, some clarifications about the ASE approach are necessary before performing the comparison. In fact, the ASE and the KJ models are based on rather different hypotheses, and slight modifications of the ASE approach are necessary:

- while the KJ model is designed to work with log-transformed variables, our method can be directly applied to the native regionalized variable (e.g. the index flood in the application of Section 3). To make the comparison more direct, here the reference variable for the ASE model is set to $\log(Q_{\text{ind}})$.

- in our approach, the selection of donor basin is based on the hierarchical organization of nested basins; in general, this introduces more than one ASE estimator for each target basin, as well as cases in which no donor basins are available because the target site is not connected to any gauged one. To make the two approaches comparable, when more than one estimator is present for the same target site, we consider only one value, taking the average of the available ASE values. If no ASE estimates are available, the regional value is adopted.

The calibration procedure for the ASE model confirmed that a threshold distance of log(10) can be considered appropriate, even when the model is applied to the logarithmic index flood. The results reported in Fig. 6(a) appear to be quite similar to those obtained for the untransformed Q_{ind}. In this plot, each point represents a single basin, different from Fig. 4 which (more generally) reports a circle for each connection $\{t,d\}$. The results obtained calibrating the KJ approach are reported in Fig. 6(b) (simplified version of the model) and in Fig. 6(c) (generalized version). The generalized version appears slightly more accurate than the simplified one.

The legends in Fig. 6 report some useful statistics, in particular the percentage of processed basins, i.e. how many regional predictions (computed during the calibration phase) are suitable to be improved. For the ASE model, it includes all the basins with $D \leq D_{\mathrm{lim}}$. For the KJ simplified version, it includes all the points having at least one neighbour within 8 km (i.e. points out of the bisector of panel (b)), while it has a trivial meaning for the KJ generalized version, since all the points are actually processed because a threshold distance does not exist. The higher percentage of basins processed by the ASE approach compared to the KJ model reflects the fact that, in this case study, the ASE method has a wider range of application. A comparison of these results also shows that our model has, on average, better performances than the

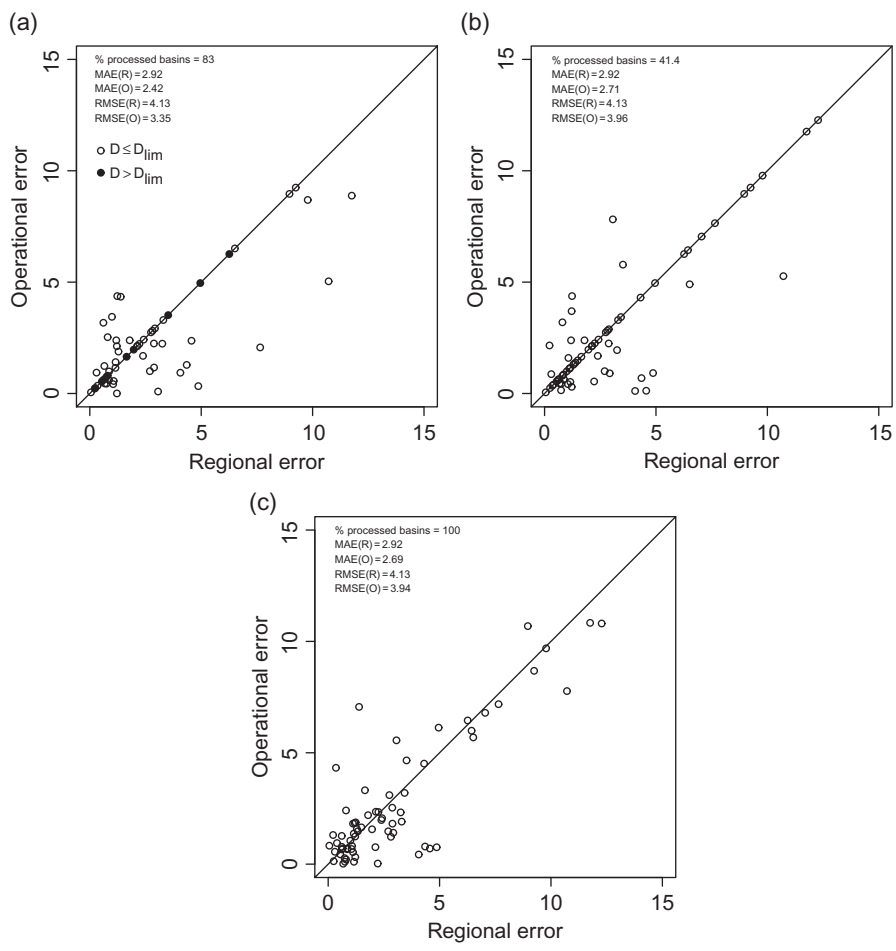


Fig. 6 Operational *vs* regional errors for the estimation of the log-transformed index flood compared: (a) ASE model with $D_{\mathrm{lim}} = \log(10)$; (b) KJ model with $\alpha_{\mathrm{KJ}} = 1$; and (c) KJ model with $\alpha_{\mathrm{KJ}} = \exp(-0.087 D_C)$.

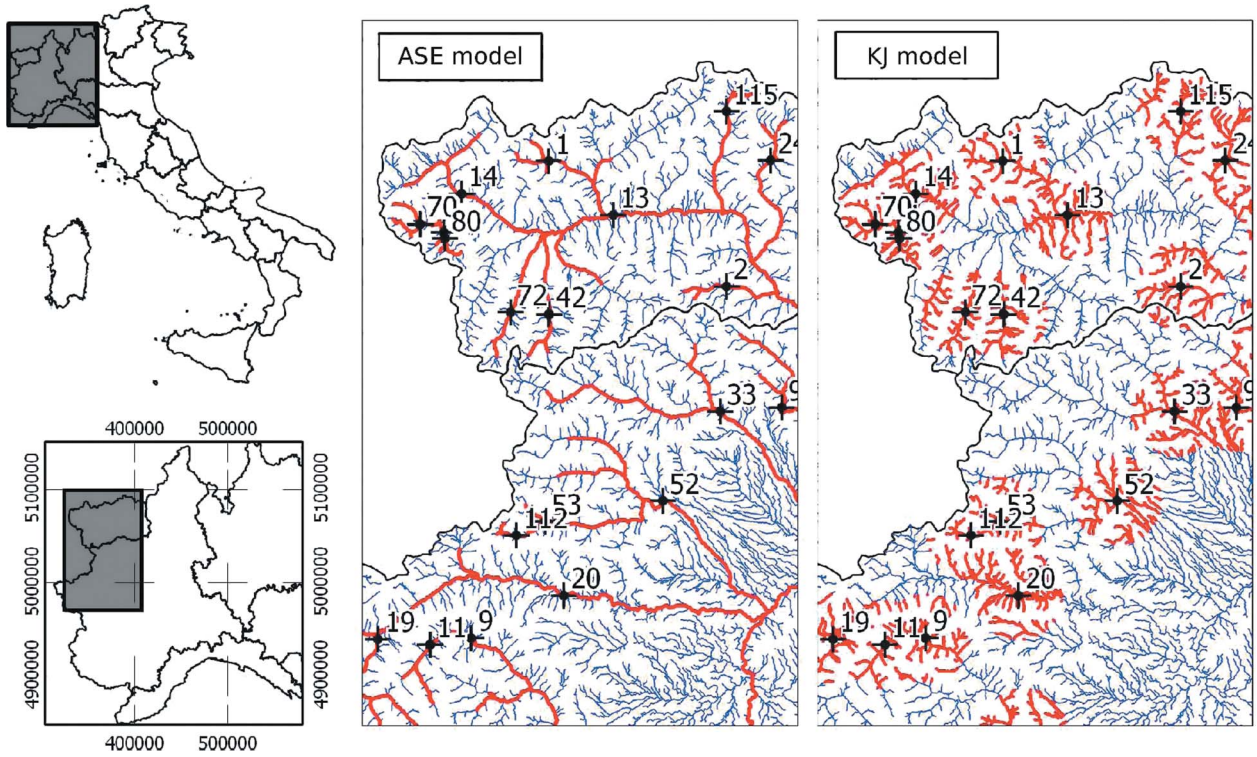


Fig. 7 Maps of applicability of ASE and KJ (simplified type) models for the case study area. The highlighted part of the drainage network represents the points where the models are applicable making use of the donor stations represented by the code numbers.

KJ approach (both versions), since the overall errors (see MAE and RMSE reported over the plots) are smaller. All the models are able to reduce the overall error with respect to the pure regional approach, as is apparent from the MAE and the RMSE (labelled R) in Fig. 6.

The very different nature and applicability of the ASE and KJ models can be examined considering the river reaches wherein the models can actually be applied. For the river network in the study region, the results are mapped in Fig. 7, where the domain of applicability is represented as a thicker line. This representation highlights the different results obtained for the propagation of information: for a highly heterogeneous area like the case study, the along-stream information propagation appears more suitable because it has a larger area of applicability.

5 DISCUSSION AND CONCLUSIONS

The Along-Stream Estimation (ASE) approach proposed herein hinges on the river structure to perform an information transfer towards ungauged basins. This integrates standard regional procedures because it is based on local relationships, as the estimation is performed considering only nested catchments. Along-stream and regional estimates can therefore be combined to develop a general framework for improved evaluation of a given hydrological variable, as well as its variance at ungauged locations.

In general, when two or more models are available for the same purpose, one can consider one of the following scenarios:

- Model competition: the results of different models (in our work "propagated" and regional predictions) can be evaluated separately and then compared, in order to identify which model is more efficient in the reconstruction of the variable of interest. In the case study presented here, propagated and regional predictions show different reliability, depending on the location of the target site and, in particular, on its distance from the donor site. From this perspective, the aim of the propagation of information is to identify an alternative procedure that is more appropriate for the analysis at some ungauged basins.
- Model cooperation: the output of one model is used to initialize the other model. In this work, for instance, the regional estimate is used as an additional parameter in the propagation function and thus contributes to the final along-stream

prediction. This approach can be interpreted as follows: the propagation of information can be used to locally improve the regional model estimate, accounting for specific information at a donor site.

- Model combination: given different estimates of the same variable, one combines them through suitable relationships aiming to minimize the variance of the resulting estimator.

The application of the ASE procedure is based on the ideas of both cooperation and competition with the regional model. In particular, the regional model tries to catch the "global" variability of hydrological variables, without considering the "local" structure of the river that can be accounted for by the propagation method. An important feature of the method, compared to other approaches for local correction, is that, even if the target site is close enough to the donor, the propagation of information is done only if the donor is suitable. This approach demonstrated its feasibility in a case study characterized by many data-scarce basins, allowing one to exploit all the available information concerning the index flood estimation.

To conclude, the along-stream approach is suitable for application wherever a regional model is available and the uncertainty of the regional predictions is provided. It exploits the local (sample) information to improve the regional estimates, and with a particular propensity for areas with short data records, because it does not require the data-demanding estimation of cross-correlation or variogram functions to represent the spatial variation of discharge.

Acknowledgements The study was supported by the Italian Ministry of Education through grant no. 2008KXN4K8. The authors acknowledge Alberto Viglione and an anonymous reviewer for their insightful comments.

REFERENCES

Claps, P. and Laio, F., 2008. Aggiornamento delle procedure di valutazione delle piene in Piemonte, con particolare riferimento ai bacini sottesi da invasi artificiali, vol. I (Updating of the procedures for flood evaluation in Piemonte, vol. I. Technical report, in Italian). Italy: Politecnico di Torino. http://www.idrologia.polito.it/piene/PienePiemonte08_Volume1.pdf.

Gottschalk, L., 1993a. Correlation and covariance of runoff. *Stochastic Hydrology and Hydraulics*, 7 (2), 85–101.

Gottschalk, L., 1993b. Interpolation of runoff applying objective methods. *Stochastic Hydrology and Hydraulics*, 7 (4), 269–281.

Gottschalk, L., *et al*., 2006. Mapping mean and variance of runoff in a river basin. *Hydrology and Earth System Sciences*, 10 (4), 469–484.

Gottschalk, L., Leblois, E., and Skoien, J. O., 2011. Correlation and covariance of runoff revisited. *Journal of Hydrology*, 398, 76–90.

Grimaldi, S.. *et al*., 2011. Statistical hydrology. *In*: P. Wilderer, ed. *The science of hydrology*, volume 2 of *Treatise on Water Science*, 479–517. Elsevier, ISBN: 978–0–444–53199–5.

Hosking, J.R.M. and Wallis, J.R., 1997. *Regional frequency analysis: An approach based on L-moments*. Cambridge: Cambridge University Press.

Institute of Hydrology, 1999. *Flood estimation handbook*. Wallingford, UK: Institute of Hydrology.

Kjeldsen, T.R. and Jones, D.A., 2007. Estimation of an index flood using data transfer in the UK. *Hydrological Sciences Journal*, 52 (1), 86–98.

Kjeldsen, T.R. and Jones, D.A., 2009. An exploratory analysis of error components in hydrological regression modeling. *Water Resources Research*, 45, W02407.

Kjeldsen, T.R. and Jones, D.A., 2010. Predicting the index flood in ungauged UK catchments: on the link between data-transfer and spatial model error structure. *Journal of Hydrology*, 387 (1–2), 1–9.

Laio, F., *et al*., 2011. Spatially smooth regional estimation of the flood frequency curve (with uncertainty). *Journal of Hydrology*, 408, 67–77.

Skoien, J.O., Merz, R., and Bloeschl, G., 2006. Top-kriging—geostatistics on stream networks. *Hydrology and Earth System Sciences*, 10 (2), 277–287.

Stedinger, J.R. and Tasker, G.D., 1985. Regional hydrologic analysis. 1. Ordinary, weighted, and generalized least-squares compared. *Water Resources Research*, 21 (9), 1421–1432.

US National Research Council, 1988. *Estimating probabilities of extreme floods*, Washington DC, USA: National Academy Press.