# 6

# CHOOSING AND ASSIMILATING FORCING DATA FOR HYDROLOGICAL PREDICTION

David C. Garen[1]

[1]*United States Department of Agriculture, Natural Resources Conservation Service, National Water and Climate Center, Portland, Oregon, USA, 97232*

## 6.1 ABSTRACT

Developing meteorological forcing data for input to hydrological models is an essential first step in modelling and prediction, whether for gauged or ungauged basins. The most common source of forcing data is meteorological stations. There are different constraints on station selection depending on the purpose of the modelling, whether the simulation is for model experimentation and testing, estimating hydrograph changes due to watershed or climate changes, or real-time streamflow forecasting. Considerations in station selection include data quality, timeliness, and spatial representativeness. Real-time forecasting poses particularly stringent requirements of station data timeliness and quality. To use station data as model input, they must be spatially interpolated over the watershed. One useful technique to do this is elevationally detrended kriging, which involves computing relationships of meteorological quantities (specifically precipitation and temperature) with elevation to describe vertical variability, subtracting this from the data to obtain residuals, then applying ordinary kriging to describe horizontal variability. The interpolation produces spatial (i.e., gridded) fields of precipitation and temperature at a daily or smaller time step, which can then be input directly to fully distributed hydrological models, or they can be averaged over the watershed or sub-areas thereof for lumped or semi-distributed models. Other interpolation techniques are usually required for other meteorological variables due to insufficient stations being available or due to the physical characteristics of the quantity not lending themselves to a kriging type of spatial interpolation (e.g. wind). Although preparation of forcing data can require significant database and

software infrastructure, especially for real-time forecasting, any hydrological modelling exercise must begin with good forcing data. In ungauged basins, without streamflow measurements to use as a check on simulation skill, it is especially critical to ensure that model forcings are accurately prepared.

## 6.2 RÉSUMÉ

La création de données de forçage météorologique comme données d'entrée dans les modèles hydrologiques constitue un premier pas essentiel dans la modélisation et la prévision, que ce soit pour les bassins jaugés ou les bassins non jaugés. Les stations météorologiques constituent la source la plus courante de données de forçage. Il existe différentes contraintes quant au choix de la station suivant le but de la modélisation, selon que la simulation soit faite à des fins d'expérimentation de modèle et d'essais, d'estimation des changements hydrographiques en raison de changements climatiques ou au niveau du bassin ou de prévision des débits en temps réel. Les facteurs à considérer dans le choix d'une station englobent la qualité des données, la rapidité de production des données et la représentativité spatiale. La prévision en temps réel pose des exigences particulièrement rigoureuses en fait de qualité et de rapidité de production des données de la station. Pour que l'on puisse se servir des données de la station en tant que données d'entrée du modèle, celles-ci doivent être interpolées spatialement à l'échelle du bassin hydrographique. Un moyen utile pour y arriver est d'employer la méthode du krigeage avec modèle de tendance, qui suppose le calcul des relations des quantités météorologiques (en particulier les précipitations et la température) avec recours à l'altitude pour décrire la variabilité verticale, en soustrayant ces valeurs des données en vue d'obtenir des données résiduelles, puis en appliquant le krigeage ordinaire pour décrire la variabilité horizontale. L'interpolation produit des champs spatiaux (c. à-d. sur une grille) de précipitations et de température à un intervalle de temps quotidien ou plus petit, qui peuvent ensuite être entrés directement dans les modèles hydrologiques entièrement distribués. Il est également possible d'en établir la moyenne en fonction de l'ensemble du bassin ou de certaines de ses sous-zones pour des modèles localisés ou semi-distribués. D'autres techniques d'interpolation sont habituellement nécessaires pour d'autres variables météorologiques en raison d'un nombre insuffisant de stations disponibles ou du fait des caractéristiques physiques de la quantité qui ne se prêtent pas à une interpolation spatiale par krigeage (p. ex. le vent). Bien que la préparation des données de forçage exige parfois une infrastructure logicielle et des bases de

données considérables, en particulier pour la prévision en temps réel, tout exercice de modélisation hydrologique doit commencer par de bonnes données de forçage. Dans les bassins non jaugés, sans mesures de débit à utiliser pour la vérification des compétences liées à la simulation, il est particulièrement essentiel de veiller à ce que les forçages de modèle soient préparés avec exactitude.

## 6.3 INTRODUCTION

An initial step of fundamental importance in hydrological prediction is developing the meteorological forcing data to be used as model input. Even if the stream to be modelled and predicted is ungauged, forcing data must still be used to define the system inputs. Without good inputs, either due to the lack of sufficient meteorological stations or due to poor processing and utilization of the station data available, one cannot expect to achieve accurate predictions. Good estimates of inputs, therefore, are essential to the success and usefulness of any system simulation exercise.

In most hydrological modelling applications, forcing data are taken from measurements at meteorological stations. While there are some examples of the use of forcings from radar, remote sensing, or atmospheric modelling (e.g., Mahfouf *et al*., 2007; Pietroniro *et al*., 2007), these are not yet common and, in some regions of high spatial variability (e.g., mountainous areas), not yet feasible. Issues relating to the use of station data in hydrological modelling, then, are of central importance. These issues include data quality, timeliness, spatial representativeness, and spatial interpolation.

These issues often do not receive thorough attention in hydrological model documentation and user manuals, giving the hydrologist rather incomplete immediately available guidance. Data quality, timeliness, and spatial representativeness are generally not addressed explicitly, presumably assuming that the hydrologist has already done a screening of stations based on these considerations and knows how to do so. Regarding spatial interpolation, sometimes models provide built-in methods for interpolating / extrapolating / averaging station data to model spatial computational units, but these tend to be very simple or based on certain assumptions about the station network that may or may not be valid (e.g., Anderson, 1973; Leavesley *et al*., 1983). For example, some built-in techniques either compute a weighted average or make a one-to-one assignment of stations to the model spatial computational units and/or require the specification of (time-invariant) elevation lapse rates (for either precipitation or temperature). Such a technique could be appropriate in a given basin, or it may be excessively rigid, incomplete, or oversimplified. Sometimes, models offer little flexibility in how the forcings are to be prepared, requiring the selection of one of the built-in methods rather than allowing the user to prepare forcings in any way desired external to the model in a preprocessing step and then supplying them to the model as input. The latter, of course, would allow the user to tailor the processing of station data into forcings for model spatial computational units, but it does put more burden on the user to have an appropriate technique at hand. In any case, the user should pay very close attention to how the station data are utilized so as to be conscious of how the forcings are prepared rather than uncritically choosing some pre-existing technique offering simply because it is convenient.

This paper presents some thoughts, ideas, considerations, and techniques for using station data in hydrological modelling and prediction. These topics apply equally to gauged and ungauged basins.

## 6.4 DATA REQUIREMENTS FOR DIFFERENT TYPES OF PREDICTION AND MODELS

Hydrological prediction can have different meanings. Three categories of what might be considered “prediction” would be: (1) Simulating the hydrograph to reproduce it as best as possible (e.g. comparing the accuracies of different models or calibrations thereof); (2) Estimating changes in the hydrograph due to past or anticipated watershed or climate changes; and (3) Real-time streamflow forecasting. All of these applications require forcing data, although there are some differences in the constraints in station usage for each application. Note that while these types of prediction are typically made in gauged basins (allowing model calibration and prediction verification), the same need exists in ungauged basins for high-quality forcings, for without this, neither of the two settings will produce successful results.

There are different types of models that can be applied for these prediction categories. The primary distinction to be made here is between statistical models and continuous process simulation models. Statistical (or empirical) models are often regression-based, such as those commonly used for long-range streamflow volume forecasts (e.g. Garen, 1992), although they could also include, for example, neural network models, where physical hydrological processes are not explicitly represented in the model structure. Process simulation models, in contrast, operate on a daily or shorter time step and have mathematical representations of a greater or lesser level of detail to represent the major hydrological water storages and fluxes that affect the flow of water into and out of the watershed.

For statistical models, the station data requirements are less stringent than for process simulation models. For the former, station data need only be good indices of the target flow to be predicted; absolute magnitudes of measured quantities do not have to be correct but only need to have a consistent relationship with the target. On the other hand, for process simulation models, the station data have to have accurate measurements in terms of absolute amounts so that the inputs to the watershed (mass and energy) are quantitatively correct. This is a much more demanding requirement than just being a consistent index.

Another difference is that for statistical models, not necessarily all stations must or even should be used. Optimization algorithms are often applied to search for combinations of predictor stations that minimize forecast error. Not all stations are necessarily required to minimize the error. In contrast, all stations, except anomalous ones with unrepresentative microclimate effects (Figure 6.1), would generally be used to define the input (e.g., precipitation, temperature) fields for process simulation models. The example shown in Figure 6.1 illustrates the importance of understanding the spatial variability of precipitation and temperature, determining if the available stations are capable of representing them, and recognizing (and perhaps excluding) stations that are not spatially representative.

An important consideration in streamflow forecasting is the real-time availability of station data. For research-mode studies, such as historical simulation or impact assessment studies, real-time station data availability is not an issue, and any stations with sufficient data can be used. This might also include discontinued stations. For forecasting, however, a more stringent data availability criterion must be applied. It does the forecaster no good to use stations in the model forcing data setup that will not be available when they are needed in forecast mode. For forecasting, then, forcing fields and model calibrations should be based only on those stations that will actually be available and usable in real-time. This places a limitation on the stations that can be selected.

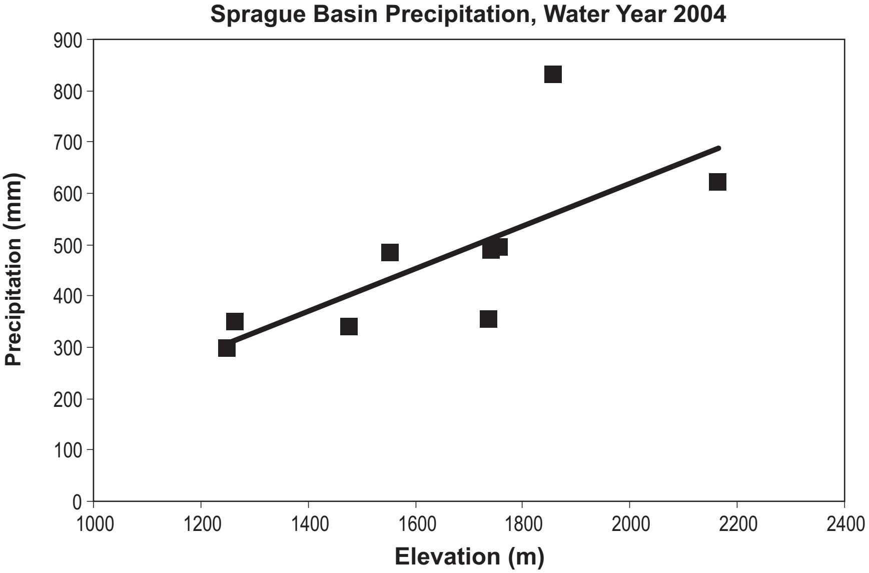

***Figure 6.1*** *Precipitation-elevation relationship for annual total precipitation in water year 2004 at meteorological stations in the watershed of the Sprague River in southern Oregon, USA. Deciding whether the anomalous station lying far above the regression line should be used for spatial interpolation of precipitation fields requires some investigation regarding the spatial representativeness of this station.*

Data quality and continuity is also important in all applications, although perhaps more critically in real-time forecasting. This issue manifests itself most commonly in missing values. It is troublesome to try to use a station in research simulation studies that has many missing values, as these must either be filled in with estimates, excluded from calculating forcing fields when missing, or the station not used at all. For real-time forecasting, these issues exist as well, but missing value detection and estimation also have to be done in real-time via an automated process for expediency and timeliness.

In fact, automated processing for input data preparation in real-time forecasting is a major requirement. Automated processing includes the following activities that must be done unattended: data retrieval from sources; data quality checks; estimation of missing data (could be optional depending on model setups); pre-processing (such as spatial interpolation); and formatting for model input. Human review of the results of this automated processing is also advisable. The rapid and automated execution of these functions is a non-trivial task requiring much database and software infrastructure.

## 6.5 SPATIAL INTERPOLATION

The use of station data for hydrological model application leads immediately to a spatial interpolation task of generalizing meteorological station data collected at a point scale to the spatial domain of a watershed. There are many ways to do this, some simple and some complex, and the technique used depends to a large degree on the number of stations available and the characteristics of the quantity being interpolated. Although some simple station weighting and averaging techniques are sometimes offered in hydrological models, more modern and complete techniques are available.

One general spatial interpolation technique that has found widespread usage in hydrology and other fields in recent years is the geostatistical procedure called kriging. Kriging is essentially a station weighting scheme. An estimate of a quantity at a spatial location is a weighted sum of the measurements at stations in its vicinity. The station weights are determined for each spatial location (most commonly grid cells in a geographic information system) in the domain to be interpolated via the kriging algorithm. The weights are a function of distance and the spatial correlation structure of the variable as represented by the semivariogram, which describes how the difference between values of the quantity at two spatial locations increases with distance between the locations (which is equivalent to, but the inverse of, a spatial correlation function, which decreases with distance). The station weights are greater for the nearest stations and smaller for the more distant stations, with the station weights summing to 1.

There are many flavours and variations of kriging, depending on specific characteristics of the data to be interpolated. One of the main issues is whether the data exhibit systematic trends in space related to a geographical characteristic, such as elevation or latitude and longitude. If this is the case, these systematic trends must either be removed from the data before applying the kriging algorithm, or the kriging framework must otherwise be designed to account for this factor affecting the spatial distribution of the quantity. Recent reviews and algorithm comparisons include Goovaerts (2000), Zhang and Srinivasan (2009), Ly *et al*. (2011), Tobin *et al*. (2011), and Feki *et al*. (2012). One such technique, elevationally detrended kriging, as applied to precipitation and temperature data is described below. This technique is highlighted here because it has been shown in the comparison studies to perform well, is conceptually straightforward, and is operationally practicable.

Garen and Marks (2005) selected this technique for use in snowpack simulations after a review of previous literature on kriging techniques.

Elevationally detrended kriging (Garen *et al*., 1994) is appropriate where elevation is the primary deterministic external factor affecting the behaviour of a meteorological variable. This is the case for precipitation, which generally increases with elevation due to orographic processes, and for temperature, which decreases with elevation. Detrended kriging divides the spatial variability of the meteorological quantity into two components: vertical and horizontal. The vertical component is described by a linear regression relationship of the quantity with elevation, which is subtracted from the data. The horizontal component is described by ordinary kriging of these detrended residuals.

The steps in the algorithm are shown in Figure 6.2. In this implementation of detrended kriging, a simplification is made by using a linear semivariogram. Doing so makes the kriging station weights invariant in time because the weights are independent of the slope and intercept of the semivariogram line. (Without this simplification, a separate semivariogram would have to be specified for each time step, greatly increasing the complexity and computational cost of the processing.) With the linear semivariogram, the kriging station weight calculation is made for all grid cells in the domain once at the beginning of the processing. From this point, the algorithm enters a loop for each time step in the time series of data to be interpolated. While a daily time step is common, shorter or longer time steps can also be accommodated in the algorithm. The calculations for each time step consist of: calculating the linear regression elevation relationship; subtracting this from the data to obtain residuals; kriging of the residuals for each grid cell in the domain; computing the deterministic elevational trend at each grid cell; and, adding the deterministic trend to the kriged residual for each grid cell to obtain the final interpolated field.

There are some implicit assumptions in this implementation. One is that the domain to be interpolated has a relatively homogeneous precipitation and temperature regime; for example, there are no strong orographic barriers within the domain that would create very different elevation relationships for different sub-areas. Another assumption is that the station density is sufficient to give a reasonable representation of the essential vertical and horizontal distribution of the precipitation and temperature fields. A final assumption is

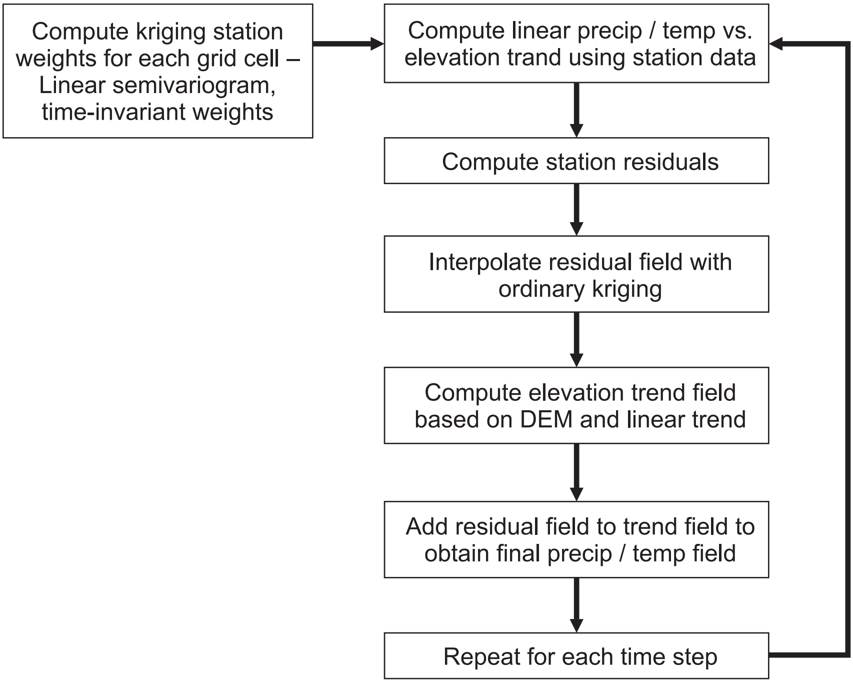
Compute kriging station weights for each grid cell – Linear semivariogram, time-invariant weights
Compute linear precip / temp vs. elevation trand using station data
Compute station residuals
Interpolate residual field with ordinary kriging
Compute elevation trend field based on DEM and linear trend
Add residual field to trend field to obtain final precip / temp field
Repeat for each time step

***Figure 6.2*** *Detrended kriging flowchart (DEM = digital elevation model).*

that the length and width of the spatial domain is moderate enough in size that the spatial correlation structure is reasonably represented by a linear semivariogram. This would imply that the domain to be interpolated should be "mesoscale" in size, perhaps on the order of 100 to 10 000 km$^2$.

Examples of interpolations of precipitation and temperature are given in Figures 3 and 4. These figures show both the elevation detrending relationship and the final interpolated field. In Figure 6.3, note that the Silver Creek site, lying well above the detrending line, exerts a significant influence on the interpolated precipitation in the northern part of the basin. Its large positive detrending residual causes grid cells in its vicinity also to have a large positive residual due to the kriging spatial interpolation, resulting in these cells also having precipitation above that estimated for their respective elevations by the detrending line. Similarly, the Gerber Reservoir and Quartz Mountain sites lie well below the detrending line, causing the kriging interpolation to calculate negative detrending residuals for grid cells in their vicinity in the southern part of the basin, and leading

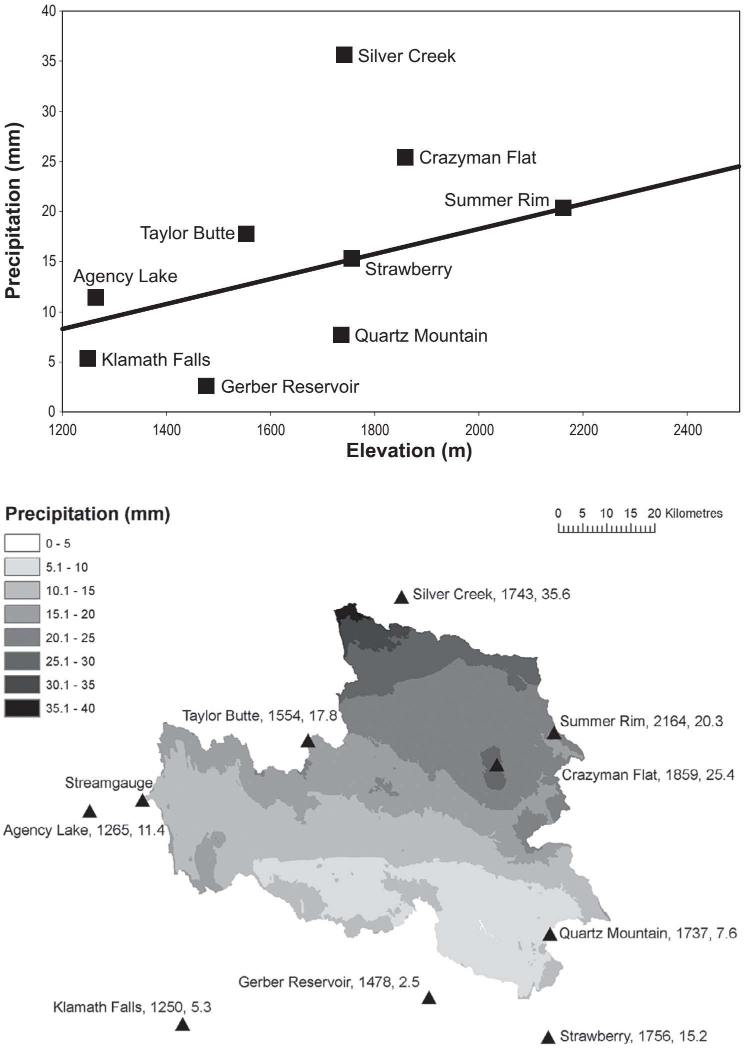

***Figure 6.3*** *Precipitation-elevation relationship and interpolated daily precipitation spatial field for 1 January 2004, Sprague River basin, southern Oregon, USA. On the map, values after the station names are, respectively, the elevation and the observed precipitation amount.*

to the final precipitation estimates being drier for their respective elevations than estimated by the detrending line. In Figure 6.4, the detrending residuals for temperature are smaller than for precipitation, so the influence of positive or negative residuals are less noticeable, and the final interpolated

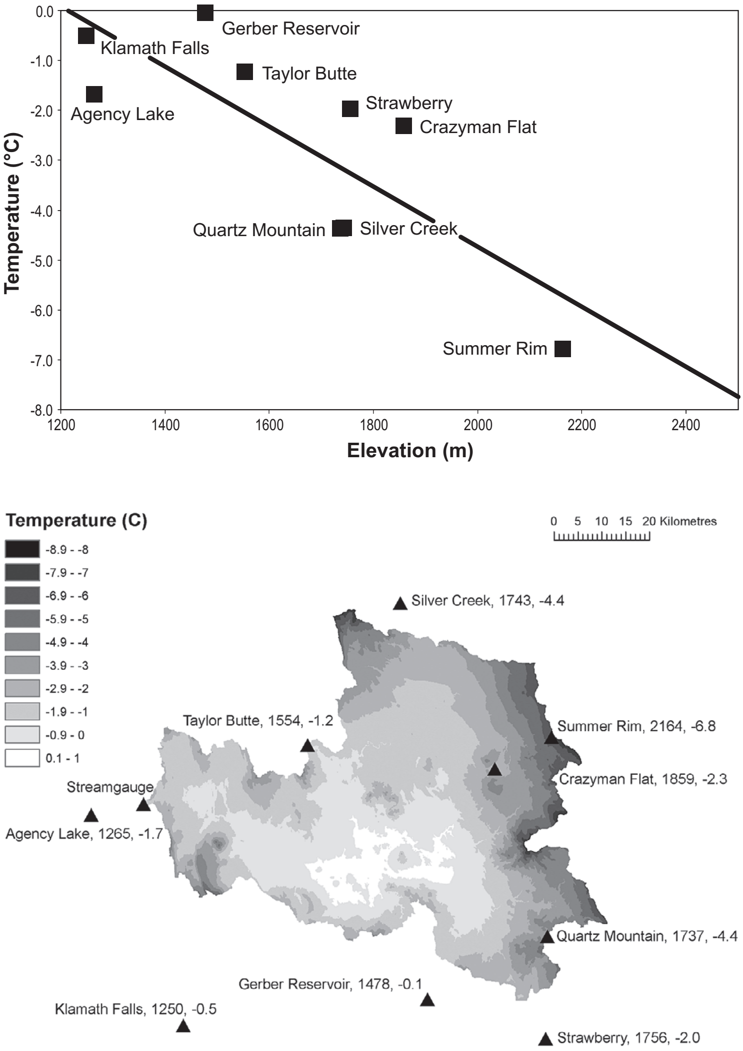

***Figure 6.4*** *Temperature-elevation relationship and interpolated temperature spatial field for the hours of 12:00-15:00 on 1 January 2004, Sprague River basin, southern Oregon, USA. On the map, values after the station names are, respectively, the elevation and the observed average temperature for the three-hour period.*

temperature follows the elevation field quite closely. Nevertheless, the residuals still have local influence, making the temperature estimates greater or less than the estimates from the detrending line for the respective grid cell elevations.

The results of such interpolations, for each time step (e.g. day) in the historical period to be simulated, can either be used directly as input to a fully distributed hydrological simulation model (requiring grid-based inputs), or the whole watershed or sub-areas thereof can be spatially averaged over the appropriate grid cells and used as input for a spatially lumped or a semi-distributed model. Note that the spatial interpolation process requires the hydrologist to consider carefully the station representativeness and data quality issues mentioned previously to ensure that the interpolation and the resulting model forcings are the best that can be done with the available information.

## 6.6 CONCLUDING REMARKS

This discussion and these examples illustrate the major considerations in selecting and interpolating data for the preparation of time series of hydrological model forcings. Careful station selection, attention to data quality, and the use of a robust and conceptually solid spatial interpolation technique are all prerequisites for a successful hydrological modelling effort.

As demonstrated, some essential meteorological station data can be spatially interpolated, but it must be remembered that the adequacy of the result is strongly dependent on station density and spatial representativeness. Precipitation and temperature are the easiest to interpolate; other meteorological variables, such as humidity, wind, and solar radiation, do not lend themselves as readily to the detrended kriging method due to sparse station density and other deterministic geographical factors for these quantities, hence other methods must be used if the model requires these additional variables (Garen and Marks, 2005). In any case, the hydrologist must establish that the station network can indeed support the preparation of adequate forcing data; if not, then there is little reason to proceed with a modelling effort, as no system can be simulated well without good estimates of the inputs.

Preparation of model forcings in a manner such as that described here gives the hydrologist confidence in the appropriateness of the system inputs given the station network and the terrain. The hydrologist can then trust the forcings and look to other model components and parameters for refining model skill. Whether in a gauged or ungauged basin, high-quality forcings are essential. Indeed, in an ungauged basin, the forcings may take on even greater importance than in a gauged basin, because there is no opportunity to use streamflow observations as a check on the adequacy of the forcings. In any case, it is evident that preparation of forcings is worth significant care and effort as the first prerequisite for successful hydrological modelling.

# REFERENCES

Anderson, E.A. 1973. National Weather Service River Forecast System – Snow accumulation and ablation model. NOAA Technical Memorandum NWS HYDRO-17, United States Department of Commerce, National Oceanic and Atmospheric Administration, National Weather Service, Silver Spring, Maryland, USA.

Feki, H., M. Slimani, and C. Cudennec. 2012. Incorporating elevation in rainfall interpolation in Tunisia using geostatistical methods. *Hydrological Sciences Journal* 57(7): 1294-1314.

Garen, D.C. 1992. Improved techniques in regression-based streamflow volume forecasting. *Journal of Water Resources Planning and Management* 118(6): 654-670.

Garen, D.C., G.L. Johnson, and C.L. Hanson. 1994. Mean areal precipitation for daily hydrologic modeling in mountainous regions. *Water Resources Bulletin* 30(3): 481-491.

Garen, D.C. and D. Marks. 2005. Spatially distributed energy balance snowmelt modelling in a mountainous river basin: Estimation of meteorological inputs and verification of model results. *Journal of Hydrology* 313: 126-153.

Goovaerts, P. 2000. Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall. *Journal of Hydrology* 228: 113-129.

Leavesley, G.H., R.W. Lichty, B.M. Troutman, and L.G. Saindon. 1983. Precipitation-Runoff Modeling System: User's manual. Water-Resources Investigations Report 83-4238, United States Geological Survey, Denver, Colorado, USA. 207 pp.

Ly, S., C. Charles, and A. Degré. 2011. Geostatistical interpolation of daily rainfall at catchment scale: The use of several variogram models in the Ourthe and Ambleve catchments, Belgium. *Hydrology and Earth System Sciences* 15: 2259-2274.

Mahfouf, J-F., B. Brasnett, and S. Gagnon. 2007. A Canadian precipitation analysis (CaPA) project: Description and preliminary results. *Atmosphere-Ocean* 45(1): 1-17.

Pietroniro, A., V. Fortin, N. Kouwen, C. Neal, R. Turcotte, B. Davison, D. Verseghy, E. D. Soulis, R. Caldwell, N. Evora, and P. Pellerin. 2007. Development of the MESH modelling system for hydrological ensemble forecasting of the Laurentian Great Lakes at the regional scale. *Hydrology and Earth System Sciences* 11:1279-1294.

Tobin, C., L. Nicotina, M. B. Parlange, A. Berne, and A. Rinaldo. 2011. Improved interpolation of meteorological forcings for hydrologic applications in a Swiss Alpine region. *Journal of Hydrology* 401: 77-89.

Zhang, X. and R. Srinivasan. 2009. GIS-based spatial precipitation estimation: A comparison of geostatistical approaches. *Journal of the American Water Resources Association* 45(4): 894-906.

## Paper citation:

Garen, D.C. 2013. Choosing and assimilating forcing data for hydrological prediction. In: *Putting Prediction in Ungauged Basins into Practice*. eds. J.W. Pomeroy, P.H. Whitfield, and C. Spence. Canadian Water Resources Association. 89-100.