YSSP Report
Young Scientists Summer Program

Process-Aware Interpolation Technique for Downscaling Hydrological Variables

Author: Marko Kallio
Email: marko.k.kallio@aalto.fi

Approved by

Supervisor: Dr. Peter Burek
Co-supervisor: Dr. Sylvia Tramberend
Program: Water
16.11.2020

This report represents the work completed by the author during the IIASA Young Scientists Summer Program (YSSP) with approval from the YSSP supervisor.

It was finished by 16th Nov 2020 and has not been altered or revised since.

This research was funded by IIASA and its National Member Organizations in Africa, the Americas, Asia, and Europe.

YSSP Reports on work of the International Institute for Applied Systems Analysis receive only limited review. Views or opinions expressed herein do not necessarily represent those of the institute, its National Member Organizations, or other organizations supporting the work.

Abstract

Water is an essential resource for human society. Numerous approaches have been developed in order to assess the availability of water for our societies. Information on the availability of water is readily available from various databases, which, however, often come in spatial aggregation which is too coarse for detailed analysis or local use cases. In this research, we propose a Process-Aware Interpolation (PAI) technique based on previous research in advanced areal interpolation. In areal interpolation, the value of a variable from a source zone is reallocated among intersecting target zones. In PAI, ancillary information on process which causes the values being interpolated is used to improve the quality of interpolation.

We test the PAI methodology in a surrogate modelling context by downscaling runoff outputs from the Community Water Model (CWatM) at 30 arc-minute resolution and compare the downscaled output to CWatM model runs at 5 arc-minute resolution. We develop two surrogate models – simplified models emulating a more complex one - based on machine learning (Random Forest Regression) and classical statistical methods (Ordinary Least Squares Regression). The surrogate models are used within the PAI framework as the ancillary information guiding the interpolation. The quality of the interpolation is assessed against a full run of CWatM at 5 arc-minute resolution, and compared to the surrogate models outside of the PAI framework as well as two simpler PAI benchmarks – a constant ancillary variable (rainfall) and an expert-knowledge based model.

We find that the developed surrogate models perform significantly better when used within the PAI framework than outside. Further, PAI with the simpler benchmarks can produce comparable quality interpolation to the PAI with surrogate models. The quality of the interpolation is, however, highly dependent on the quality of the source data.

1 Introduction

The availability of timely and adequate quality data about water resources are crucial for efficient management of the water resource. Ideally, data should be collected at the location-of-interest, but environmental monitoring networks are sparse and such locally collected information may not be available. Researchers and practitioners have long turned to environmental and hydrological modelling to overcome the lack of observations at a certain location. Complex distributed environmental models, however, can be challenging to learn and operate, and depending on the type of the model, can be time-consuming to run [1,2].

On the other hand, a large number of environmental datasets are freely available from various databases; for instance Lindersson et al.[3] list and describe 124 free and open datasets useful in (socio-)hydrology. The global datasets, however, often come in mismatching spatial, or temporal aggregation which renders them suboptimal, particularly for local applications. Areal interpolation methods have been developed to assess the spatial aggregation issue by reallocating data from source zones (areal units) to overlapping target zones [4–6]. Simple areal interpolation – reallocation of source zone values based on the proportional overlapping areas with target zones – is a standard procedure in hydrology. Researchers working with population dynamics and small area estimation have further developed the method allowing the use of ancillary information in the interpolation step [4–7] in Dasymetric Mapping (DM). In DM, the estimates from simple areal interpolation are refined using quantified ancillary data. Further, Tobler [8] developed Pycnophylactic Interpolation (PP) method to estimate the distribution of a variable within specified zones. Only a few studies have explored the advanced areal interpolation methods in hydrology: Kallio et al. [9,10] used them to downscale runoff, and Chen et al. [11] refined groundwater extraction estimates using DM.

In this study, we extend the approach of Kallio et al. [9,10] to describe a Process-Aware Interpolation (PAI) method for downscaling hydrological variables. In PAI, DM is replaced with a Dasymetric *Modelling* (dasymetric modelling and mapping are used interchangeably hereafter) [12], where the ancillary information is provided by a *model* describing the process governs the variable under interpolation. In this context, the choice of the model is flexible and the aim of it is to reproduce the dynamical behaviour of the interpolated variable within source zones.

We explore the PAI methodology through an experimental case study of the Upper Bhima Basin in Southern India. Community Water Model (CWatM) [13], a full physically based distributed hydrological model, is run in 30 arc-minute (0.5°; approximately 55 km at the equation) and in 5 arc-minute resolution. We downscale runoff output from CWatM$_{30min}$ to 5 arc-minute resolution using the PAI framework, and evaluate the quality of downscaling against the output of CWatM$_{5min}$. This places our experiment in the context of surrogate modelling – simple(r) approximation of an expensive physical model.

We develop a Random Forest Regression and Ordinary Least Squares surrogate models for CWatM and use them in PAI as the ancillary model. We further use to alternative ancillary models as points of comparison: non-dynamic, constant mean precipitation and an expert-knowledge based model.

The following sections are organised as follows: in Section 2 we provide a brief descriptions to the areal interpolation methods used in PAI as well as an introduction to the case study. In Section 3 we provide the preliminary results in testing the performance of the PAI methodology, and Section 4 gives our summary conclusions.

2 Materials and Methods

In this section we first introduce the areal interpolation concepts in Section 2.1. The following Section 2.2 gives a description of the case study. For a more thorough introduction to areal interpolation and it's applications we recommend Comber and Zheng [6].

2.1 Areal interpolation methods

2.1.1 Area Weighted Interpolation (AW)

In areal interpolation, unknown values at *target* zones are estimated based on values associated with overlapping *source* zone(s).[4–6] At it's simplest, the estimation is done based on the proportional overlapping areas (Area Weighted interpolation; AW),

$$\hat{V}_t = \sum_s^{s\cap t} V_s \frac{A_{t\cap s}}{A_s} \qquad (1)$$

where $\hat{V}_t$ is the estimated value V at target zone t, V_s is the known value at source zone s, and A is area. AW preserves mass/volume in the values of the source zones, if the source zone is entirely covered by target zones. Thus, if V is *extensive* (i.e. density), it should first be converted to a count statistics (e.g. precipitation data expressed as depth per time unit, mm day^{-1}, should be converted to volume, e.g. m^3 day^{-1}) prior to interpolation.

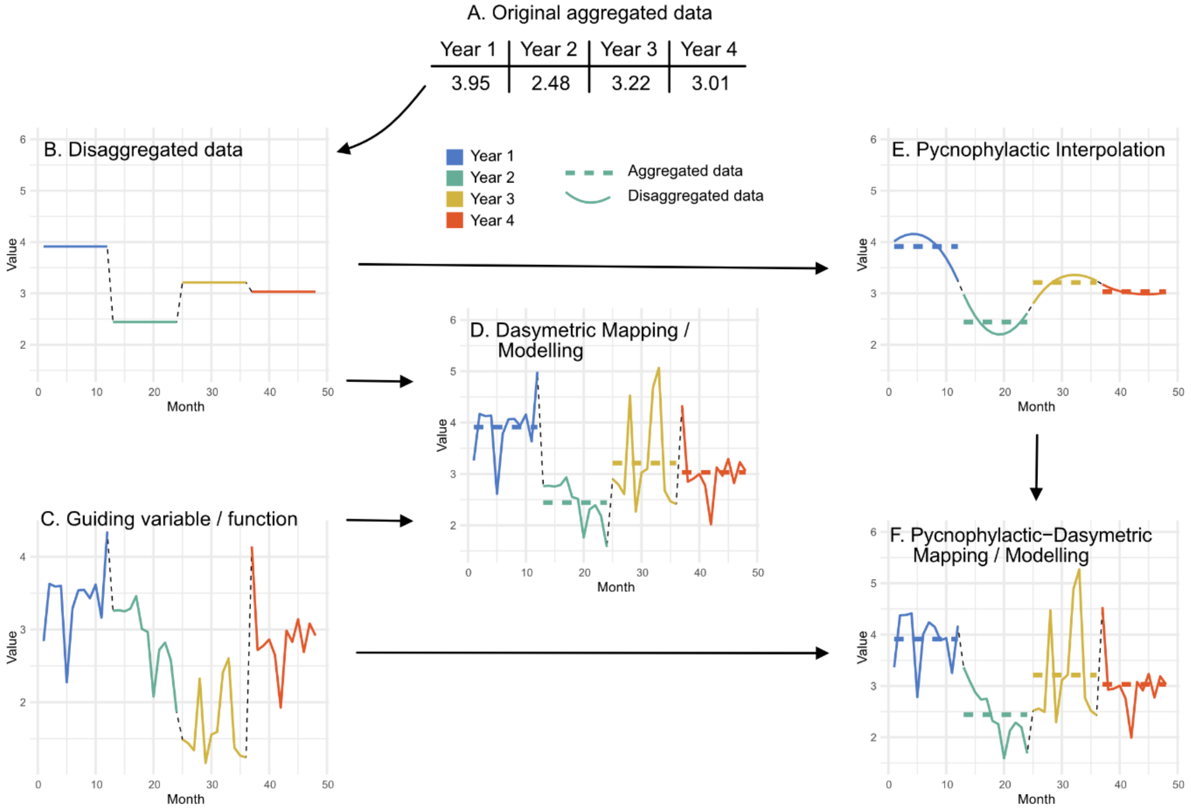

Figure 1. One dimensional analogues for the advanced areal interpolation methods discussed in this study. Minimum required data are A) aggregated data, and C) a guiding variable or model output. B) shows disaggregation akin to Area Weighted Interpolation, while D) shows output using Dasymetric Mapping approach, E) showcases Pycnophylactic Interpolation, and F) a combined Pycnophylactic-Dasymetric Mapping.

Figure 1B shows a 1-dimensional example of AW, where an aggregated annual values are distributed to the 12 months of the year. Each target unit (month) receives the value of their parent because there is no temporal overlap over the source units (year). Figure 2B on the other hand shows an example in two dimensions, where the source data is interpolated to non-conforming target zones. Here, the target areas which transcend the source zone boundary have a different value compared to the target zones entirely within the source zone.

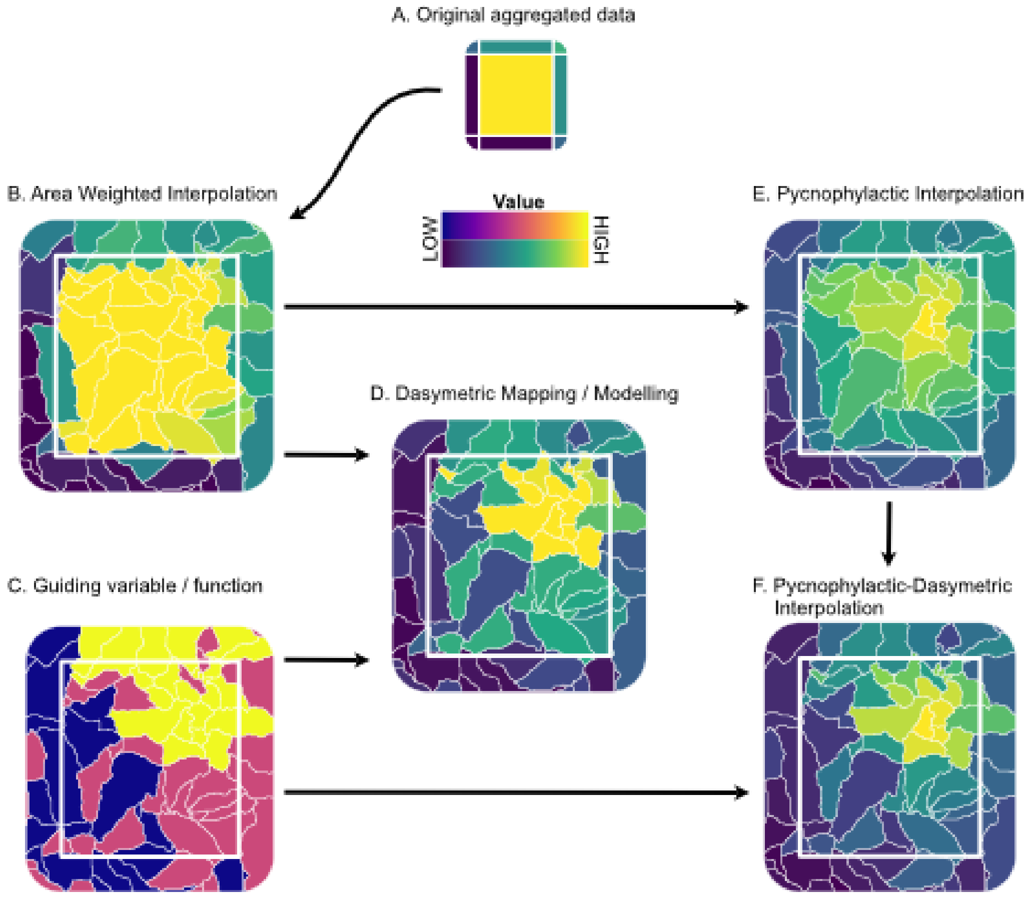
A. Original aggregated data
B. Area Weighted Interpolation
Value
LOW
HIGH
E. Pycnophylactic Interpolation
D. Dasymetric Mapping / Modelling
C. Guiding variable / function
F. Pycnophylactic-Dasymetric Interpolation

Figure 2. A schematic spatial example visualizing the discussed methods. A) is showing the aggregated data in source zones while C) gives the non-conforming target zones and the ancillary variable for them. B) shows the output using Area Weighted Interpolation. Output of Dasymetric Mapping is given in D), and E) shows the output of Pycnophylactic Interpolation. F) gives the Pycnophylactic-Dasymetric Interpolation. Adapted from Kallio et al. [9]

2.1.2. Dasymetric Mapping (DM) and Modelling (DM)

DM modifies the approach of AW by introducing an ancillary variable to guide distributing V_s to target zones,

$$\hat{V}_t = \sum_s^{s\cap t} V_s A_{s\cap t} \frac{A_{t\cap s} X_t}{\sum_t^{t\cap s} A_{t\cap s} X_t} \qquad (2)$$

where X_t is the (finite and positive numberic) value of ancillary variable in target zone t. X_t can be replaced with a model $f(X_t)$ in DM, where the aim of the model is to describe the distribution of V in target zones t. A common approach is to estimate the relationship between X_t and V_s using a linear model, or to employ expert knowledge[14]. Many variants of the DM have been developed, for instance

Mennis [15] extends the dasymetric approach to spatio-temporal domain, and Nagle et al [12] develop a maximum entropy modelling-based approach able to account for uncertainties in the source and ancillary data. In this study we, however, concentrate on the simple DM shown in Eq. 2.
The approach is visualized in Figure 1D, reduced to one dimension, and Figure 2D shows a visualization with spatial data for which the approach is originally developed. Comparing to AW, the inclusion of an ancillary variable enables estimation of the distribution of the interpolated quantity also for those target zones which are entirely within a single source zone.

2.1.3 Pyncophylactic Interpolation (PP)

Pycnophylactic Interpolation [8] (PP) is another type of areal interpolation method, which attempts to estimate the internal distribution of V_s within the source zone *s*. Tobler's original study [8] describes PP for gridded data, where the source zones are disaggregated using a higher-resolution grid, and a smooth surface is fitted so that the mass or volume of V_s is preserved within the source zones. Rase [16] describes PP for Triangulated Irregular Networks, while Kallio et al. [10] adapts PP for use with arbitrarily shaped polygon networks. Examples are shown for one dimension in Figure 1E and for spatial data in Figure 2E.

2.1.4 Combined Pycnophylactic-Dasymetric Interpolation (PP-DM)

Kallio et al. [9] combine the two areal interpolation approaches, PP and DM, into a combined Pycnophylactic-Dasymetric Interpolation (PP-DM). In this approach, rather than using AW as the basis for adjusting the distribution, PP-DM first estimates a smooth surface via PP, which' values are then adjusted based on the ancillary variable(s), or a model. A visualization example can be seen in Figure 1F and Figure 2F for one and two dimensional cases, respectively. It is noteworthy that the result is highly similar to DM when the smooth surface does not deviate much from AW (years 3 and 4 in Figure 1F), but significant difference can be seen where this deviation is large (end of year 1, start of year 2).

2.2 Case study method

In order to explore the capabilities of PAI, we perform a downscaling experiment for the Upper Bhima Basin in India in a surrogate modelling context. Community Water Model (CWatM) [13] outputs from a global modelling run at 30 arc-minute resolution are downscaled to 5 arc-minute grid (see Figure 3). We develop a surrogate ,pfrö for CWatM runoff output based on Random Forest (RF) , and use the outputs of the RF model at 5 minute resolution as the ancillary model in PAI. To evaluate the performance of PAI, it is compared to the output of CWatM run in the 5 arc-minute resolution, as well as to the RF surrogate model applied at the higher resolution. The following sections detail the case study area and data, and the surrogate models in more detail, followed by our choice of performance metrics.

2.2.1 Upper Bhima Basin

The study area is located in Southern India in the state of Maharashtra. The Bhima River is one of the main tributaries of the Krishna River flowing to the Indian Ocean. The basin has highly variable topography and rainfall distribution; to the west is the Ghats mountains which receives more than 4000 mm rainfall per year, and is the source of the Bhima River. To the east is the plains of Deccan Plateau, which receives less than 500 mm rainfall per year. [17] The area is influenced by the Monsoon climate, leading to 80-90% of annual rainfall falling in the wet season from June to October. It is a heavily cultivated area with over 70% of land area under agricultural use [17]. The study area topography is shown in Figure 1 along with the runoff outputs from CWatM in both resolutions. The distribution of rainfall is nearly identical (Pearson correlation = 0.99) to the runoff shown in Figure 3B,C.

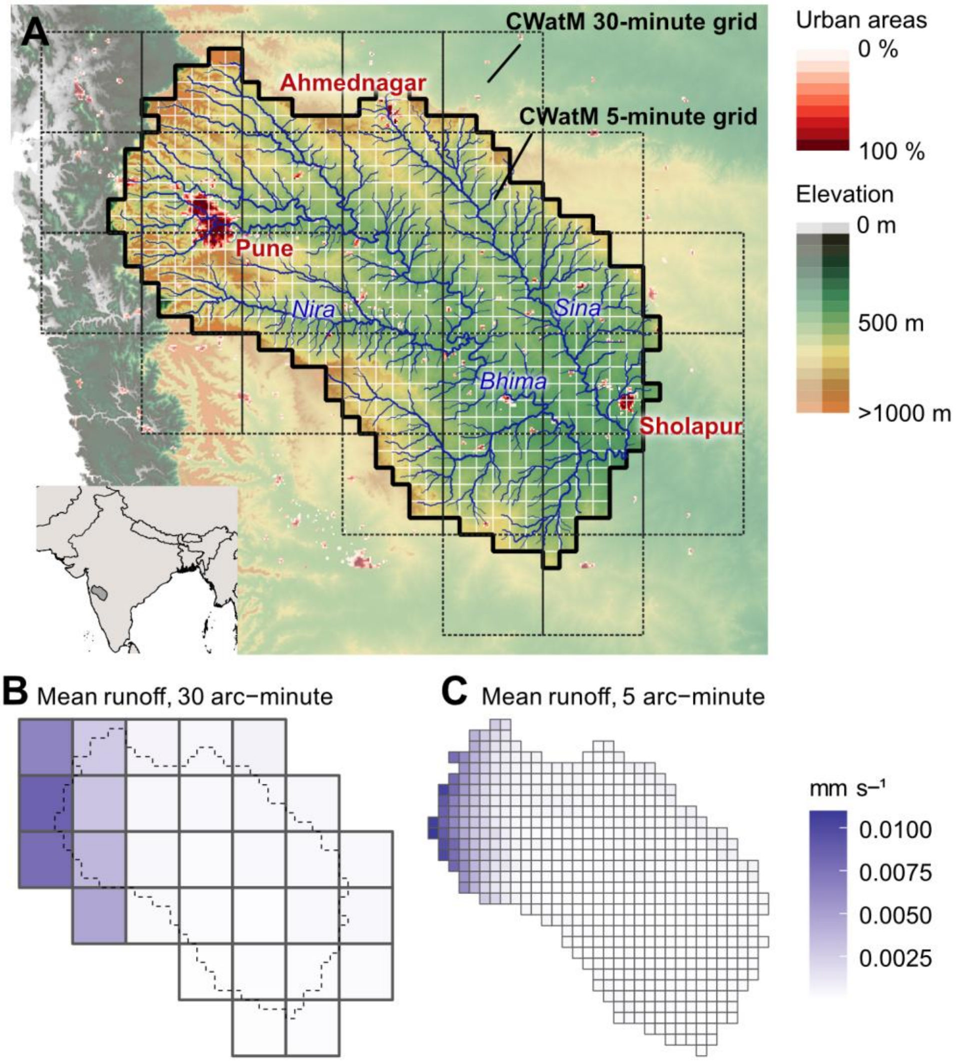

Figure 3. A) Upper Bhima Basin study area showing the CWatM model grids used in the study, B) mean runoff output between 2001 and 2010 from CWatM$_{30min}$, and C) mean runoff output from CWatM$_{5min}$ for the same period. In the case study, runoff from B) is downscaled with an aim to achieve the distribution shown in C).

2.2.2 Surrogate models and comparison

In response to the challenges posed by distributed environmental models, surrogate modelling can be employed. Surrogate models mimic a full-fledged high-fidelity (fidelity in this context refers to degree of realism) model by simplified representation [1,2] A surrogate model can thus be termed as a metamodel, or a model-of-model. Razavi et al. [2] divide surrogate modelling approaches to two broad categories: 1) response surface modelling, where a data-driven process is used to emulate the output of the original high-fidelity model, and 2) lower-fidelity modelling, where some components of the original model are left out, or simplified. For a more detailed treatise on surrogate modelling in hydrological context, we refer the reader to Asher et al [1], and Razavi et al [2].

The surrogate modelling approach we use in this study is the function approximation approach, described in Razavi et al. [2], in which the surrogate model is trained using previously evaluated original model at a number of data points. Kriging interpolation is a common method used in relation to water resources [2], making other interpolation methods, like PAI, a viable alternative. We develop two surrogate models for CWatM based on Random Forest Regression [18] (RF) and Ordinary Least Squares Regression (LM). Both surrogate models are constructed by training and testing the model predicting runoff using 30 arc-minute resolution CWatM run. The model specification for both RF and LM is given in Equation 3;

$$Runoff = P_t + P_{t-1} + P_{t-2} + P_{t-3} + DSP + SM + \ Elevation + DUNE \qquad (3)$$

where *P* is precipitation, *t* is timestep, *DSP* stands for the number of Days Since previous Precipitation event, *SM* is soil moisture, and *DUNE* is the topographic index dissipation per unit length [19]. DUNE was derived from the 15 arc-second HydroSHEDS [20] data, and both elevation and DUNE were aggregated to the CWatM resolutions. The lagged precipitation and DSP were derived from the daily precipitation data (see Table 1). Both, RF and LM, surrogate models perform well with the test data (30 arc-minutes; 27360 data points), as shown in Table 2. Because the ancillary variable in DM needs to be a finite positive value, the predictions made by RF or LM models need to be processed so that any negative values are eliminated. Here we choose to shift the RF and LM results so that $Runoff = Runoff + \min(Runoff)$. Both of the surrogate models are used in the PAI framework as the process model in DM and PP-DM.

Table 1. Input data used in the case study.

Variable		Spatial Resolution	Data source
Precipitation (P)			
	Daily P	5min, 30min	Input data for CWatM [21–24]
	Previous day P	5min, 30min	
	P two days ago	5min, 30min	
	P three days ago	5min, 30min	
	Number of days since last P event	5min, 30min	
Soil Moisture (SM)			CWatM
	Daily SM	5min, 30min	
Elevation			HydroSHEDS [20]
	Mean elevation	15s	
	Mean DUNE	15s	
Runoff			CWatM
	Daily runoff	5min, 30min	

We further use two benchmarking process representations: a constant ancillary variable (CA) mean precipitation in the target zone over the entire study period 2001-2010, and a model based on expert knowledge, shown in Equation 4:

$$Runoff = aP + \frac{b}{DUNE} + \frac{Elevation}{c} \qquad (4)$$

where *a = 5000, b = 10, and c = 20* are coefficients which were arbitrarily chosen by the author of this study. The aim of the expert model in the context of PAI is not to model the actual runoff values, but to estimate a map of potential runoff distribution which can be used within the PAI methodology to guide the interpolation.

2.2.3 Performance evaluation

The performance of PAI is examined through four metrics:

1. Root Mean Square of Error (NRMSE) normalized to the mean of observations,
2. Percent Bias (PBIAS),
3. Pearson correlation coefficient (r), and
4. Kling-Gupta Efficiency (KGE) [25].

We find a 7.9 % positive bias in the CWatM$_{30min}$ runs compared to CWatM$_{5min}$. Since the PAI methods described here are pycnophylactic – i.e. they preserve the interpolated quantity, it leads to a maximum theoretically attainable KGE being 0.921 in the case when *r* = 1, and α = 1 in Equation 4,

$$KGE = \sqrt{(r-1)^2 + (\alpha-1)^2 + (\beta-1)^2} \tag{4}$$

where r is the Pearson correlation coefficient, α is a measure of variability error $^{\sigma_{sim}}/_{\sigma_{obs}}$, and β is the percent bias $^{\mu_{sim}}/_{\mu_{obs}}$. σ stands for the standard deviation, and μ for the mean. Thus, KGE is an aggregated metric which covers bias, variability and dynamics. To further evaluate the quality of interpolation, we compute local KGE to gain additional insight on the spatial variablility of the performance.

Table 2. Performance of the surrogate models (RF, LM) on the 30 minute resolution, their application to a 5 minute grid, and the performance of PAI methods (AI, DM, PP, PP-DM). The dasymetric approaches reported are temporally constant precipication (CA), Random Forest surrogate (RF), OLS surrogate (LM), and an expert knowledge-based model (expert). Metrics highlighted in bold perform as well as, or better than, either of the two surrogate models.

			Performance			
	Model	Set	NRMSE %	PBIAS %	r	KGE
	RF	30min, test set	21.9	-0.1	0.98	0.94
	LM	30min, test set	29.4	-0.4	0.96	0.93
	RF	5min, benchmark	53.4	-2.7	0.90	0.58
	LM	5min, benchmark	58.6	-33.1	0.82	0.61
1	AI	5min	58.6	8.0	0.82	**0.71**
2	DM (CA)	5min	**41.6**	7.9	**0.91**	**0.80**
3	DM (RF)	5min	**44.7**	7.9	**0.90**	**0.82**
4	DM (LM)	5min	**42.0**	7.9	**0.91**	**0.81**
5	DM (expert)	5min	**44.7**	7.9	**0.90**	**0.77**
6	PP	5min	**47.8**	7.9	0.88	**0.77**
7	PP-DM (CA)	5min	**41.9**	7.9	**0.91**	**0.85**
8	PP-DM (RF)	5min	**47.7**	7.8	0.88	**0.85**
9	PP-DM (LM)	5min	**43.8**	7.8	**0.90**	**0.85**
10	PP-DM (expert)	5min	**42.1**	7.9	**0.91**	**0.83**

3 Results and discussion

We find that the global performance of the surrogate models against CWatM$_{5min}$ are significantly worse than the performance on the CWatM$_{30min}$ test set (Table 2) with the error growing to twice as large. The surrogate models also have larger bias, particularly with the LM surrogate, which grows from -0.4 % to -33.1%. This is largely because the linear model predicts negative runoff values, thus skewing the distribution of values. Correlation coefficient for both surrogate models, however, is good being 0.9 for RF and 0.82 for LM.

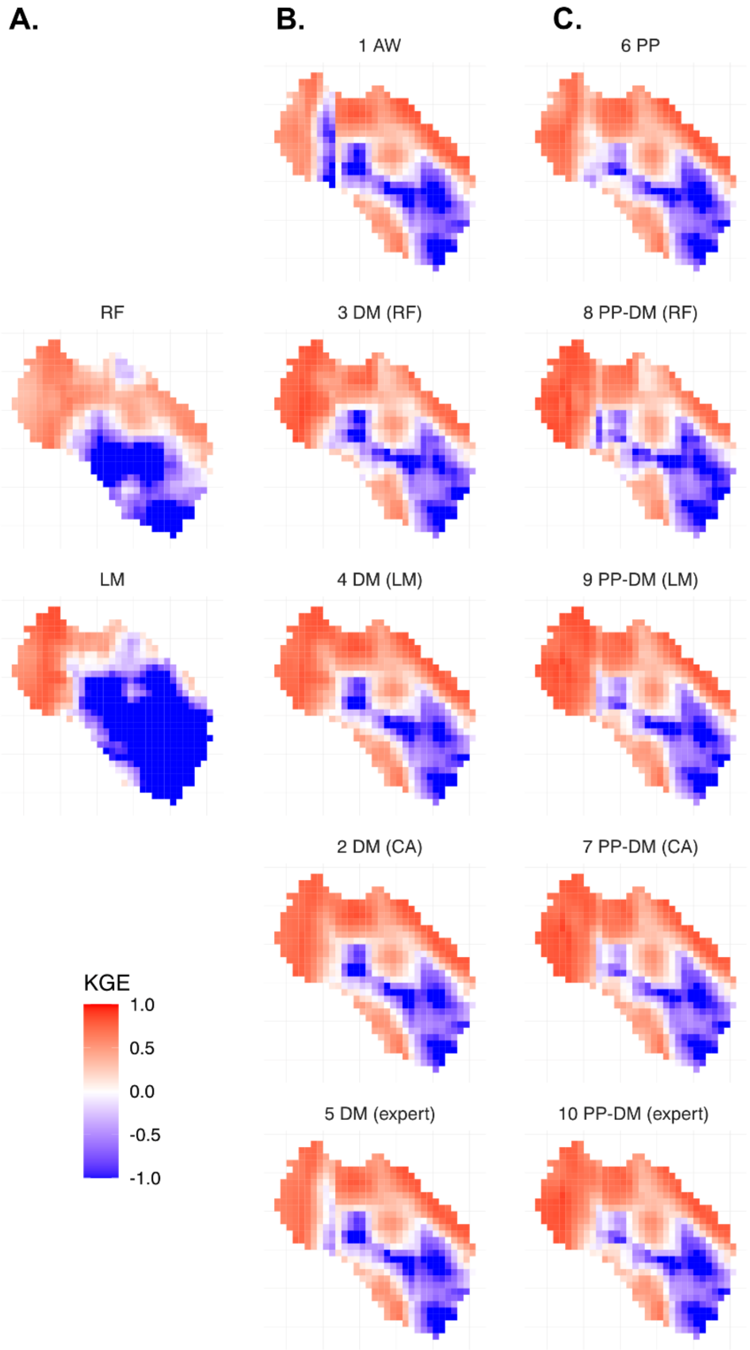

Figure 4. Local R^2 of the PAI estimated runoff to the output of CWatM$_{5min}$ within a rolling 5x5 grid cell window. Column A) shows the surrogates, column B) the PAI methods based on Area Weighted Interpolation, and column C) PAI methods based on Pycnophylactic Interpolation.

It is particularly noteworthy that the surrogate model performance is enhanced when used with the PAI framework; error is reduced, correlation is similar, and KGE is considerably improved. Change in bias, however is not dependent on the particular PAI method or surrogate model; rather, it is determined by the bias in source data, which is not modified by the PAI framework based on redistribution of values.

While the global performance metrics give favourable results for the PAI, the performance varies considerably in different parts of the Upper Bhima Basin. Figure 4 shows that KGE of all PAI methods perform well in the south-eastern, eastern and northern part of the basin, while performance is particularly poor in the central and south-western part. This is primarily due to the bias between the source (CWatM$_{30min}$) and the target (CWatM$_{5min}$) data, but differences in the hydrological conditions between these areas may also play part. This is evident when examining AW in Figure 4B shows the correlation of CWatM$_{30min}$ data against CWatM$_{5min}$ leading to a conclusion that the source runoff data is a poor approximation of the target data in the areas where PAI methods perform poorly. We attribute this to the different meteorological forcing data the two different resolution CWatM runs use. It is, however, noteworthy that the PAI methods are able to attain relatively good KGE values in parts of the poorly correlated areas visible in Figure 4B (AW). Particularly encouraging for the PAI methodology is that the performance of the surrogate models is improved across the entire case study area. In fact, even the simple expert-knowledge based ancillary model performs better than either of the two surrogate models in most of the study area.

Surrogate modelling is commonly used in order to reduce the computation times of complex models. We measured the computation times of our software implementation written in R language [26]. In the case study area consisting of 28 source zones, 567 target zones and 3652 timesteps (10 years with daily timestep), AW and DM run through under 6 seconds with an Intel Core i7-6600 processor, while PP runs in approximately 5 minutes. As a comparison, CWatM for the study area, over 10 years runs through in the same 5 arc-minute resolution takes approximately an hour. The current implementation based on native R code is therefore fast enough for the common motivation of reducing computation time in surrogate modelling. The developed technique has an additional surrogate-related benefit of considerably smaller high-resolution data procurement requirement than the full original CWatM model run needs (see the manual at https://cwatm.iiasa.ac.at/index.html). Further, the study stands as a proof-of-concept that the PAI methodology could be used in a surrogate modelling context.

4 Conclusions

Previous studies [9,10] dealing with hydrological variables in an advanced areal interpolation context have been conducted using constant ancillary variables. In this study, we extended this methodology to utilize spatio-temporal process models instead of constant ancillary information to improve the performance of the interpolations.

The developed Process-Aware Interpolation were tested in a surrogate modelling case study in the Upper Bhima Basin in India where we approximate Community Water Model outputs at 5 arc-minute resolution by downscaling the same model outputs at 30 arc-minute resolution using the PAI methodology. Two surrogate models (Random Forest and Ordinary Least Squares) were trained with the 30 arc-minute CWatM input and output data, and evaluated in 5 arc-minute resolution. The two surrogates were used as the process model in PAI together with two more simple process representations; constant mean precipitation and an expert-knowledge based model.

We find that the PAI methodology utilizing any of the four process representations can significantly increase the performance of runoff approximation compared to the two surrogates used outside PAI framework. However, it is evident that the methodology is highly sensitive to the quality of source data, and due to the mass preserving property of areal interpolation methods, the method is not able to correct poorly fitting source data unless more advanced versions of DM are used. While the method shows promise, properties of different choices of process modelling should be further explored. A software implementation of the developed PAI interpolation methods is under development as an R package called *dasymetric*.

References

1. Asher, M. J., Croke, B. F. W., Jakeman, A. J. & Peeters, L. J. M. A review of surrogate models and their application to groundwater modeling. *Water Resour. Res.* **51**, 5957–5973 (2015).
2. Razavi, S., Tolson, B. A. & Burn, D. H. Review of surrogate modeling in water resources. *Water Resour. Res.* **48**, (2012).
3. Lindersson, S., Brandimarte, L., Mård, J. & Baldassarre, G. D. A review of freely accessible global datasets for the study of floods, droughts and their interactions with human societies. *WIREs Water* **7**, e1424 (2020).
4. Goodchild, M. F. & Lam, N. S. N. Areal interpolation: A variant of the traditional spatial problem. *Geo-Process.* **1**, 297–312 (1980).
5. Eicher, C. L. & Brewer, C. A. Dasymetric Mapping and Areal Interpolation: Implementation and Evaluation. *Cartogr. Geogr. Inf. Sci.* **28**, 125–138 (2001).
6. Comber, A. & Zeng, W. Spatial interpolation using areal features: A review of methods and opportunities using new forms of data with coded illustrations. *Geogr. Compass* **13**, e12465 (2019).
7. Goodchild, M. F., Anselin, L. & Deichmann, U. A Framework for the Areal Interpolation of Socioeconomic Data. *Environ. Plan. Econ. Space* **25**, 383–397 (1993).
8. Tobler, W. R. Smooth Pycnophylactic Interpolation for Geographical Regions. *J. Am. Stat. Assoc.* **74**, 519–530 (1979).
9. Kallio, M., Virkki, V., Guillaume, J. H. A. & van Dijk, A. I. J. M. Downscaling runoff products using areal interpolation: a combined pycnophylactic-dasymetric method. in *El Sawah, S. (ed.) MODSIM2019, 23rd International Congress on Modelling and Simulation.* (Modelling and Simulation Society of Australia and New Zealand, 2019). doi:10.36334/modsim.2019.K8.kallio.

10. Kallio, M., Guillaume, J. H. A., Virkki, V., Kummu, M. & Virrantaus, K. Hydrostreamer v1.0 - improved streamflow predictions for local applications from an ensemble of downscaled global runoff products. *Geosci. Model Dev. Discuss.* 1–37 (2020) doi:https://doi.org/10.5194/gmd-2020-276.

11. Chen, J., Broussard, W. P., Borrok, D. M. & Speyrer, F. B. A GIS-Based Framework to Identify Opportunities to Use Surface Water to Offset Groundwater Withdrawals. *Water Resour. Manag.* doi:10.1007/s11269-019-02298-5.

12. Nagle, N. N., Buttenfield, B. P., Leyk, S. & Speilman, S. Dasymetric Modeling and Uncertainty. *Ann. Assoc. Am. Geogr. Assoc. Am. Geogr.* **104**, 80–95 (2014).

13. Burek, P. *et al.* Development of the Community Water Model (CWatM v1.04) A high-resolution hydrological model for global and regional assessment of integrated water resources management. *Geosci. Model Dev. Discuss.* 1–49 (2019) doi:https://doi.org/10.5194/gmd-2019-214.

14. Mennis, J. Dasymetric Mapping for Estimating Population in Small Areas. *Geogr. Compass* **3**, 727–745 (2009).

15. Mennis, J. Dasymetric Spatiotemporal Interpolation. *Prof. Geogr.* **68**, 92–102 (2016).

16. Rase, W.-D. Volume-preserving interpolation of a smooth surface from polygon-related data. *J. Geogr. Syst.* **3**, 199–213 (2001).

17. Surinaidu, L., Bacon, C. G. D. & Pavelic, P. Agricultural groundwater management in the Upper Bhima Basin, India: current status and future scenarios. *Hydrol. Earth Syst. Sci.* **17**, 507–517 (2013).

18. Breiman, L. Random Forests. *Mach. Learn.* **45**, 5–32 (2001).

19. Loritz, R. *et al.* A topographic index explaining hydrological similarity by accounting for the joint controls of runoff formation. *Hydrol. Earth Syst. Sci.* **23**, 3807–3821 (2019).

20. Lehner, B., Verdin, K. & Jarvis, A. New Global Hydrography Derived From Spaceborne Elevation Data. *Eos Trans. Am. Geophys. Union* **89**, 93–94 (2008).

21. Pai, D. S. *et al.* Development of a new high spatial resolution (0.25° × 0.25°) long period (1901-2010) daily gridded rainfall data set over India and its comparison with existing data sets over the region. *MAUSAM* **65**, 1–18 (2014).

22. Cucchi, M. *et al.* WFDE5: bias-adjusted ERA5 reanalysis data for impact studies. *Earth Syst. Sci. Data* **12**, 2097–2120 (2020).

23. Weedon, G. P. *et al.* The WFDEI meteorological forcing data set: WATCH Forcing Data methodology applied to ERA-Interim reanalysis data. *Water Resour. Res.* **50**, 7505–7514 (2014).

24. Burek, P. *et al.* Development of the Community Water Model (CWatM v1.04) – a high-resolution hydrological model for global and regional assessment of integrated water resources management. *Geosci. Model Dev.* **13**, 3267–3298 (2020).

25. Gupta, H. V., Kling, H., Yilmaz, K. K. & Martinez, G. F. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *J. Hydrol.* **377**, 80–91 (2009).

26. R Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, 2019).