HYDROINFORMATICS AND DIVERSITY IN HYDROLOGICAL ENSEMBLE PREDICTION SYSTEMS

Thèse

Darwin Brochero

Doctorat en génie civil et génie des eaux
Philosophiæ doctor (Ph.D.)

Québec, Canada

Résumé

Nous abordons la prévision probabiliste des débits à partir de deux perspectives basées sur la complémentarité de multiples modèles hydrologiques (diversité). La première exploite une méthodologie hybride basée sur l'évaluation de plusieurs modèles hydrologiques globaux et d'outils d'apprentissage automatique pour la sélection optimale des prédicteurs, alors que la seconde fait recourt à la construction d'ensembles de réseaux de neurones en forçant la diversité.

Cette thèse repose sur le concept de la diversité pour développer des méthodologies différentes autour de deux problèmes pouvant être considérés comme complémentaires. La première approche a pour objet la simplification d'un système complexe de prévisions hydrologiques d'ensemble (dont l'acronyme anglais est HEPS) qui dispose de 800 scénarios quotidiens, correspondant à la combinaison d'un modèle de 50 prédictions météorologiques probabilistes et de 16 modèles hydrologiques globaux.

Pour la simplification, nous avons exploré quatre techniques: la *Linear Correlation Elimination*, la *Mutual Information*, la *Backward Greedy Selection* et le *Nondominated Sorting Genetic Algorithm II* (NSGA-II). Nous avons plus particulièrement développé la notion de participation optimale des modèles hydrologiques qui nous renseigne sur le nombre de membres météorologiques représentatifs à utiliser pour chacun des modèles hydrologiques.

La seconde approche consiste principalement en la sélection stratifiée des données qui sont à la base de l'élaboration d'un ensemble de réseaux de neurones qui agissent comme autant de prédicteurs. Ainsi, chacun d'entre eux est entraîné avec des entrées tirées de l'application d'une sélection de variables pour différents échantillons stratifiés. Pour cela, nous utilisons la base de données du deuxième et troisième ateliers du projet international MOdel Parameter Estimation eXperiment (MOPEX).

En résumé, nous démontrons par ces deux approches que la diversité implicite est efficace dans la configuration d'un HEPS de haute performance.

Abstract

In this thesis, we tackle the problem of streamflow probabilistic forecasting from two different perspectives based on multiple hydrological models collaboration (diversity). The first one favours a hybrid approach for the evaluation of multiple global hydrological models and tools of machine learning for predictors selection, while the second one constructs Artificial Neural Network (ANN) ensembles, forcing diversity within.

This thesis is based on the concept of diversity for developing different methodologies around two complementary problems. The first one focused on simplifying, via members selection, a complex Hydrological Ensemble Prediction System (HEPS) that has 800 daily forecast scenarios originating from the combination of 50 meteorological precipitation members and 16 global hydrological models.

We explore in depth four techniques: Linear Correlation Elimination, Mutual Information, Backward Greedy Selection, and Nondominated Sorting Genetic Algorithm II (NSGA-II). We propose the optimal hydrological model participation concept that identifies the number of meteorological representative members to propagate into each hydrological model in the simplified HEPS scheme.

The second problem consists in the stratified selection of data patterns that are used for training an ANN ensemble or stack. For instance, taken from the database of the second and third MOdel Parameter Estimation eXperiment (MOPEX) workshops, we promoted an ANN prediction stack in which each predictor is trained on input spaces defined by the Input Variable Selection application on different stratified sub-samples.

In summary, we demonstrated that implicit diversity in the configuration of a HEPS is efficient in the search for a HEPS of high performance.

Notation

t	Time-step
N	Number of pairs observations-forecasts
D	Total number of hydrological members in the forecast ensembles
M	Total number of m intervals to analyze the reliability diagram
c	Identification of the rank or class to analyze the uniformity in the rank histogram
o^t	Observed flow at the time t
$\mathbf{y}^t$	Ensemble flow forecast at the time t
y_i^t	i^{th} flow forecast member in $\mathbf{y}^t$
$\mathbf{Y}$	Ensemble flow forecast from $t=1$ to N
$\mathbf{o}$	Observations vector from $t=1$ to N
F	Cumulative distribution function
f	Probability density function
ϕ	Normalized variables for probability density function
Φ	Normalized variables for cumulative distribution function
$\bar{o}_m$	Conditional probability of the event as a function of the interval I_m assigned to the forecast $m \rightarrow P(o^t \mid I_m)$
r^t	Binary indicator, 1 if the event occurs for the t^{th} forecast-event pair, 0 if it does not
S_c	Number of elements of the c^{th} interval of the rank histogram ($c=1$, ..., $d+1$)
$\underset{t=1}{\overset{N}{\text{med}}}$	Median value evaluated from $t=1$ to N
μ_t	Mean ensemble flow forecasts at the time t
σ_t^2	Variance ensemble flow forecasts at the time t
χ_{t}	Estimation set
χ_{v}	Validation set
χ_{p}	Test or publication set
$\{a^t\}_{t=1}^N$	Set of a with index t ranging from 1 to N

$\underset{\theta}{\operatorname{argmin}} \; g(x\|\theta)$	The argument θ for which g has its minimum value
$\mathcal{E}(\theta\|\chi)$	Error function with parameters θ on the sample χ
w_{cp}	Weights of the components of the combined criterion (CC)
$\text{iter}_{\text{xp}}^{\mathbf{y}_i}$	Iteration number at which was eliminated the $\mathbf{y}_i$ hydrological member during the selection process in the xp experiment
$\overline{R}(\mathbf{y}_i)$	Mean rank of elimination of the $\mathbf{y}_i$ hydrological member
s	Final selection of the nm best hydrological members in the selection process
x^t	Model inputs at time t
$E()$	Expectation operator

Acronyms

AdaBoost Adaptive Boosting. 128, 142

AMALGAM A Multi-ALgorithm, Genetically Adaptive Multiobjective. 47

ANN Artificial Neural Network. v, 1–3, 9–12, 14–16, 31, 32, 34, 35, 37–41, 89, 107–111, 114–118, 120–134, 137, 138, 140–143, 147–152

BGS Backward Greedy Selection. 16, 39, 40, 45, 54, 55, 57–59, 61, 63, 64, 66, 68, 69, 73, 74, 76–82, 85–87, 89, 90, 94, 96–98, 100–102, 147, 148, 150, 151

BMA Bayesian Model Average. 46, 47

BP Back-propagation. 37

CC Combined Criterion. 55, 56, 58, 61–64, 66–68, 70, 73, 76, 77, 79, 91, 92, 97, 147, 151

CDF Cumulative Distribution Function. 25, 26

CMIM Conditional Mutual Information Maximization. 91, 92

CRPS Continuous Ranked Probability Score. 25, 26, 47, 50, 52, 55, 56, 59, 60, 62–66, 68, 70, 76–78, 87, 134, 140–142, 147–150

CV Cross-Validation. 73, 74, 76–82, 147, 150, 151

DISSENT Dynamic Input Spaces imposed by Stratified Examples propagated on artificial Neural networks Training. 130, 131, 133–142, 148, 150, 151

DREAM DiffeRential Evolution Adaptive Metropolis. 46

EA Evolutionary Algorithms. 2, 38, 40, 71

ECMWF European Centre for Medium-range Weather Forecasts. 13, 15, 45, 47–49, 53, 58, 68, 71, 73, 76, 78, 85, 86, 88, 147

EPS Ensemble Prediction System. 4, 15, 19, 23–25, 30, 45, 47–49, 53, 58, 68, 71, 73, 76, 78, 85, 86, 88, 147

ESN Echo State Network. 115, 120, 121, 125, 141

FCP Flood Contingency Plan. 7–9

FFNN Feed-Forward Neural Network. 1, 14–16, 31, 37, 107, 109, 114–118, 120, 141, 142, 151

FGS Forward Greedy Selection. 32, 39, 40, 90, 91, 116, 129, 151

FOU First-Order Utility. 91, 92

FTH Forecast Time Horizon. 3, 7, 14, 16, 49–53, 60, 62, 63, 66, 67, 69, 73, 74, 77–86, 89, 95, 98, 100, 102

GA Genetic Algorithm. 40

GLUE Generalized Likelihood Uncertainty Estimation. 10, 46

HEPEX Hydrologic Ensemble Prediction EXperiment. 4

HEPS Hydrological Ensemble Prediction System. iii, v, 1, 4–15, 30–32, 35, 39–41, 45–47, 49–53, 55–60, 66, 68–70, 73, 76–78, 80, 81, 84–88, 94–97, 100–103, 107, 120, 127, 129, 141, 147–152

HMP Hydrological Models Participation. 16, 45, 51–54, 60, 66, 68, 73, 74, 77–86, 88, 95, 97, 100–102, 147, 148, 150

HWL High Warning Level. 7, 9

IGNS IGNorance Score. 26, 27, 47, 50, 52, 55, 56, 59, 60, 62–66, 68, 70, 76–78, 83, 85, 87, 89, 93, 95–98, 100, 134, 140–142, 147–150

iqr Interquartile range. 49, 51, 61, 79, 136, 137

IVS Input Variable Selection. 16, 32, 37, 41, 107, 118, 127, 130–132, 134, 135, 138, 142, 147, 148

JMI Joint Mutual Information. 91, 92

LCE Linear Correlation Elimination. 85, 86, 89, 90, 100, 102, 148

MAE Mean Absolute Error. 25, 47, 59, 78, 116, 139–141

MCMC Markov Chain Monte Carlo. 46

MCS Multiple Classifier System. 6, 12

MDCV MeDian of the Coefficients of Variation. 50, 52, 55, 56, 61–68, 70, 76, 77, 80, 149

MEPS Meteorological Ensemble Prediction System. 9, 31, 35, 45–47, 58, 68, 70, 71, 80, 82

MI Mutual Information. 85, 86, 89, 100, 102, 148, 151

MIFS Mutual Information based Feature Selection. 91, 92

MIM Pure Mutual Information Maximization. 91, 92

MLD Multi-Level Diversity. 5, 6, 9, 11–16, 31, 35, 126, 129, 130

MOPEX MOdel Parameter Estimation eXperiment. iii, v, 15, 107, 118, 120, 129, 148

MRMR Maximum-Relevance Minimum-Redundancy. 91, 92

MSC Meteorological Service of Canada. 58, 71

MSE Mean Square Error. 19–21, 30, 35, 38, 55, 59, 75, 80, 88, 89, 116, 122, 124, 125, 133, 137, 139, 141, 148

NCDC National Climate Data Center. 120

NCEP US National Centers for Environmental Prediction. 71

NCEP/NCAR National Center for Environmental Predictions/National Center for Atmospheric Research. 120

NOAA National Oceanic and Atmospheric Administration. 120

NS Normalized Sum. 58, 61–63, 65–69, 76, 77, 79, 81–83

NSE Nash-Sutcliffe Efficiency. 116, 118, 121, 124, 125, 137, 139

NSGA-II Nondominated Sorting Genetic Algorithm II. iii, v, 16, 40, 85, 86, 97, 98, 100–102, 148, 151

NWS National Weather Service. 120

PDF Probability Density Function. 4, 5, 7, 19, 24–27, 46, 61, 87, 91, 95

PEARP Prévision d'Ensemble Action de Recherche Petite échelle grande échelle. 47

PI Persistence Index. 116–118, 122, 124, 125, 137, 139, 151

R100P Ensemble of 30 FFNNs trained with early stopping using a **R**andom sampling of **100** **P**ercent of the available information and a single predefined set of inputs variables. 14, 107, 117, 118, 120–122, 124–127, 134, 136, 138–142, 148

RD Reliability Diagram. 28, 60, 85, 148, 149

RD_{MSE} MSE assessed from the differences between conditional observed frequencies and evaluated probability thresholds in the RD. 28, 49–52, 56, 62–66, 68, 76–78, 88, 89, 93, 95–98, 100, 134, 141, 148

RH Rank Histogram. 29, 60

RTALC Revised Technological Adoption Life Cycle. 11

SAFRAN Système d'Analyse Fournissant des Renseignements Atmosphériques à la Neige. 48

SCE Shuffled Complex Evolution. 2

SIM Coupling of **S**AFRAN (Système d'Analyse Fournissant des Renseignements Atmosphériques à la Neige), **I**SBA (Interactions between Soil, Biosphere, and Atmosphere), and **M**ODCOU (MODélisation COUplée) models. 47

SOM Self-Organizing Map. 108

TIGGE **T**he Observing System Research and Predictability Experiment (THORPEX) **I**nteractive **G**rand **G**lobal **E**nsemble. 46, 58

USGS US Geological Survey. 119, 120

UTC Coordinated Universal Time. 48

Contents

Notation vii

Acronyms ix

Contents xiii

List of Tables xv

List of Figures xvii

Introduction 1
- I.1 Hydrologic modelling . . . 1
- I.2 Error and uncertainty . . . 2
- I.3 Hydrological ensemble prediction systems . . . 4
- I.4 Multi-level diversity model . . . 5
- I.5 Importance of HEPS . . . 6
- I.6 Thesis main topics . . . 8
- I.7 Hypothesis and objectives . . . 12
- I.8 Thesis structure . . . 15

I Basic Concepts 17

1 Ensemble Prediction System Evaluation 19
- 1.1 Bias, variance, and covariance . . . 19
- 1.2 Ensemble forecasts quality . . . 23
- 1.3 Verification statistics for ensemble forecasts . . . 25
- 1.4 Conclusion . . . 30

2 Machine Learning: Some Concepts and Tools 31
- 2.1 Generalization ability . . . 32
- 2.2 Clustering . . . 35
- 2.3 Artificial neural networks . . . 37
- 2.4 Feature selection . . . 37
- 2.5 Conclusion . . . 41

II HEPS Simplification **43**

3 Optimization Criteria **45**
3.1 Review of HEPS simplification . . . 45
3.2 HEPS of reference . . . 47
3.3 Hydrological models participation . . . 52
3.4 Estimation of hydrological models participation . . . 53
3.5 Results and analysis . . . 59
3.6 Conclusion . . . 69

4 Generalization in Time and Space **71**
4.1 Generalization test methodology . . . 71
4.2 Results and discussion . . . 74
4.3 Conclusion . . . 82

5 Comparison of Techniques in a General Framework of Selection **83**
5.1 General framework of the simplification scheme . . . 84
5.2 Methodology for the simplification techniques comparison . . . 84
5.3 Results and Analysis . . . 92
5.4 Conclusion . . . 99

III ANN Ensembles as HEPS **103**

6 Diversity from Dataset and Parametric Levels **105**
6.1 Stratification concept for ANN training . . . 105
6.2 Methodology . . . 107
6.3 Study area . . . 116
6.4 Results and discussion . . . 118
6.5 Conclusion . . . 123

7 Diversity from Dataset, Parametric, and Model Inputs Levels **125**
7.1 Introduction . . . 125
7.2 Methodology . . . 128
7.3 Results and discussion . . . 132
7.4 Conclusion and future work . . . 139

IV Conclusion, Contributions, and Future Work **141**

A Publications Resulting from this Thesis **149**

B MSE Decomposition - Deterministic Case **151**

C MSE Decomposition - Expected Square Error **153**

D MSE Decomposition - Multimodel Approach **155**

Bibliography **157**

List of Tables

I.1 Examples of hypothetical HEPS complexity. . . . 10

2.1 Differences of datasets nomenclature. . . . 35

3.1 Hydrological models. . . . 48
3.2 Main characteristics of the studied catchments. . . . 50
3.3 Performance of the 16-member HEPS and the 800-member HEPS. . . . 51
3.4 Hypothetical example to show the HMP concept. . . . 52
3.5 Median of 200 random selections in catchment H36. . . . 62
3.6 Results of BGS in basin H36. . . . 63
3.7 30-member HEPS scheme based on different scores. . . . 65
3.8 Selection of 100 hydrological members. . . . 67

4.1 Selection of 50 members based on the CC. . . . 75
4.2 Test based on the NS in new catchments. . . . 80

5.1 Parametrization proposed to evaluate the mutual information. . . . 89
5.2 NSGA-II set-up. . . . 91
5.3 Probabilistic performance for the 9^{th} FTH in different schemes. . . . 93
5.4 Scores for the 9^{th} FTH and 48-member HEPS for the BGS and NSGA-II. . . . 98

6.1 Stratification example. . . . 111
6.2 Example of fold data distribution. . . . 112
6.3 Neural network set-up. . . . 115
6.4 Main characteristics of the studied catchments. . . . 117
6.5 Mean MSE gain of resampling techniques. . . . 122

7.1 List of model input candidates. . . . 130
7.2 Number of input subspaces found in 30 DISSENT experiments. . . . 133
7.3 Deterministic functions to evaluate the DISSENT relevance. . . . 137
7.4 Probabilistic scores to evaluate the DISSENT relevance. . . . 139

List of Figures

I.1 Some concepts about HEPS. . . . 4
I.2 Uncertainty Cascade Model. . . . 6
I.3 Multi-Level Diversity model. . . . 7
I.4 HEPS as basis of a Decision Support System. . . . 8
I.5 Revised Technological Adoption Life Cycle (RTALC). . . . 11
I.6 Members selection scheme. . . . 13

1.1 The dartboard and the bias-variance analogy. . . . 21
1.2 Probabilistic forecasting evaluation. . . . 23
1.3 Continuous ranked probability score evaluation. . . . 26
1.4 Ignorance score evaluation. . . . 27
1.5 Reliability diagram. . . . 28
1.6 Rank histogram. . . . 29

2.1 Generalization problem. . . . 33

3.1 Catchments location. . . . 49
3.2 iqr of RD_{MSE} and δ ratio of two HEPS schemes. . . . 50
3.3 Evaluation of HMP. . . . 54
3.4 Comparison between the 800-member HEPS and 30-member HEPS. . . . 60
3.5 Evolution of the NS in terms of gain index. . . . 61
3.6 Evolution of the gain index for each score under different criteria. . . . 66
3.7 BGS and box-plots in 200 random experiments of 50 members. . . . 68

4.1 Generalization test methodology. . . . 72
4.2 Evolution of the NS to evaluate the local sensibility. . . . 77
4.3 800-member and 50-member HEPS comparison for the 9^{th} FTH 78
4.4 Evolution of the NS to evaluate the regional sensibility. . . . 79
4.5 HMP and models rank index. . . . 81

5.1 HEPS simplification based on clustering and different HMP. . . . 85
5.2 Behaviour of scores in the BGS on training and validation datasets. . . . 94
5.3 Evaluation of different selections with NSGA-II. . . . 96
5.4 Comparison of different HEPS simplification schemes of 48 members. . . . 97
5.5 Comparison between HMP results of BGS and NSGA-II. . . . 99

6.1 Assignment of similar data points to different folds. . . . 109
6.2 Flowchart of the proposed stratification methodology. . . . 110
6.3 Space to stratify. . . . 111

6.4 Basins and hydroclimatological regimes. 117
6.5 NSE performance in test for several ANN and the R100P model. 119
6.6 Train and test datasets properties. 120
6.7 Stratification results. 121
6.8 MSE normalizations and best individual stratification schemes. 122

7.1 Dynamic IVS procedure. 129
7.2 DISSENT procedure example. 132
7.3 Frequency of variable selection. 134
7.4 Interquartile range and median of the ANN ensemble errors. 135
7.5 Scatter plot of ensemble streamflow prediction and observed streamflow. 136
7.6 Reliability diagrams evaluated in R100P and DISSENT models. 137
7.7 Relative rank histograms evaluated in R100P and DISSENT models. 138

To Karol, light of my life.
To Celeste and Violeta, colours
of my universe.

Acknowledgements

I would like to express my gratitude to all my supervisors, colleagues, friends, family, who have helped and supported me throughout my thesis.

I had the good fortune to get a pair of great supervisors, Dr. François Anctil and Dr. Christian Gagné, who make a top team to integrate hydrology with machine learning concepts. To each of them I offer my heartfelt gratitude for constantly encouraging me with patience and knowledge.

To Jasper Vrugt and Stefan Schroedl, thank you for sharing the source codes of NSGA-II and Features Selection with Mutual Information, respectively. I am also indebted to the numerous tools and systems made by the Open Source community.

To Jenny Brochero and Annette Schwerdtfeger, thank you for the proofreading of this thesis.

Thank you also to Geneviève Pelletier, Marc Parizeau, and Paulin Coulibaly for their careful revision, which helped improve it.

I am very thankful to the Department of Civil Engineering and Water Resources Engineering of Université Laval. They provided the support I needed to produce and complete my thesis.

I am grateful to the Computer Vision and Systems Laboratory at Université Laval that provided an excellent environment for my research. I have been surrounded by wonderful colleagues, thank you for helping to develop the ideas in this thesis.

My work benefited of financial support from NSERC (Canada), ICETEX (Colombia), and across to databases from CEMAGREF, MOPEX project, and ECWMF.

This thesis would also not be possible without the love and support of my friends, who gave me a home away from home.

Finally I'd like to thank my wonderful wife Karol, without her support and love I'd be lost. And I'd also like to thank my great daughters, Celeste and Violeta, who keep life interesting.

Introduction

Streamflow prediction (hereafter also referred as hydrological forecasting) and its applications are numerous, but its assessment remains complex. We address this problem with two subjects extensively developed in the last decade: Hydrological Ensemble Prediction System (HEPS), i.e. systems based on multiple prediction scenarios, and the application of machine learning tools in the water science context, which is known as hydroinformatics.

So, initially we show hydrological simulation types emphasizing the conceptualization of error and uncertainty in the prediction. Consequently, we expose the nature of the HEPS as a way of dealing with the various sources of error and/or uncertainty. Subsequently, we highlight the concept of complementarity of predictors called diversity, which is strongly linked to the implementation of schemes that force variability in the forecast.

Then, we present a hypothetical example showing the operational advantages of the HEPS. Finally, we show the importance of the two issues addressed in this thesis: the simplification of a complex HEPS from the selection of predictors and the production of a HEPS with nonlinear regression models known as Artificial Neural Networks (ANNs), specifically the Feed-Forward Neural Network (FFNN) structure.

I.1 Hydrologic modelling

To understand the complexity of the streamflow prediction problem, we start with the basic outline of the hydrological simulation whose dynamics may best be understood as a cause-effect relationship (rainfall-runoff). So, we must abstract four sub-process within the catchment*:

- Spatial and temporal distribution of liquid precipitation and melt-water.
- The mechanisms of interaction between atmospheric variables and soil characteristics, especially evapotranspiration, heterogeneity of cover types, and soil layers.
- Evaluating soil moisture, hydraulic capacity, infiltration, and percolation to finally determine the drainage pathways that result in surface runoff, inter-flow, and base flow.
- Hydraulic routing scheme of different flow pathways.

*We use interchangeably catchment or basin to define an area of land where surface water converges to a single (exit) point.

This complex natural system is represented with a hydrological model. In the last decade, it has been suggested to couple models specialized for each physical sub-process, leading to the union of atmospheric models and soil-vegetation atmospheric transfer schemes [93, 119, 169]. However, by simplicity and information availability, a "standalone" hydrological model is frequently used. Depending on the type of modelling, it can be classified as:

- **Physical or white-box models**: all sub-processes are conceptualized based on the laws of physics, considering the energy and mass balance of the system. Ideally all equations have a physical meaning and parameters can be calculated from measurements of the system.
- **Data-driven or black-box models**: the dynamics and information of all sub-processes are not explicitly considered in the model. Instead, it is assumed that the data series contain the information needed to model the system without considering specific physical sub-processes. In this regard, the hydroinformatics community is recognized for their adaptation in water science of techniques such as ANN, Evolutionary Algorithms (EA), or other modern technologies for the purposes of satisfying social requirements [4].
- **Conceptual or gray-box models**: the hydrologic sub-processes are considered from semi-empirical and simplified equations of physical origin. The parameters used in these models mostly come from calibration with optimization techniques such as EA [58] or the Shuffled Complex Evolution (SCE) or one of its variants [55, 153, 154].

Hydrological models can be distinguished by their spatial discretization. Global or lumped models consider the entire basin as one unit. Distributed models subdivide the basin into grid-cells to simulate the flow pathways with cell-transmission information.

I.2 Error and uncertainty

We cannot lose sight that each hydrological model is a simplified representation of reality, which leads to an inherent error between observed and simulated streamflows. So, the aim of the modelling is to minimize such error, for which optimization techniques are used to calibrate, in some cases, dozens of parameters to adjust the response of the model to the observations. At this point, it is important to note that we always run the risk that the model may succeed as mathematical marionettes, dancing to match the observations even if their underlying premises are unrealistic [86].

Now then, based on the acceptance of the modelling error, it may be conceived as the result of the joint uncertainty regarding measurements and conceptualization. However, there are factors that may increase the uncertainty, for example when observations are not available at every point in the basin or cannot be measured to an infinite degree of precision, when initial states are unknown, and when the structure of the model does not fully capture the many processes within the basin. Additionally, the problem of scale in hydrology increases

uncertainty, since some processes that are important at a particular scale may not necessarily be at another one [138].

In the case of **streamflow prediction** based on physical or conceptual models, an additionnal source of uncertainty emerges: the prediction of atmospheric and/or meteorological variables that serve as inputs to the hydrological model. In this regard several authors have already highlighted this other source of uncertainty as the most uncertain component in the prediction process [83, 115, 145].

It is not necessary to link black-box models to a meteorological prediction system to predict streamflow, because the former usually exploits a relationship between past observations and prediction horizon. Consequently, the analysis basically involves a regression model based on hydroclimatological lagged variables. However, the selection of modelling data, the estimation of the input variables, the choice of the mathematical model and its parameters are active sources of uncertainty in this type of modelling for which ANNs are recognized for their accuracy and high computational efficiency [97].

Despite the philosophical differences between the different types of modelling, some of the greatest advances in hydrology can be expected from joint work exploring many approaches [138]. For example, it is well known that many operational hydrological systems require forecasts for several days in advance, i.e. different Forecast Time Horizons (FTHs), while prediction with black-box models is often confined to a few days of anticipation since the variable with the greatest information is generally the same unknown lagged streamflow. Prediction with black-box models may thus offer more options if it is included into a hybrid framework, taking advantage and enriching conventional prediction systems based on the propagation of precipitation forecasts into the physical or conceptual hydrological simulation models.

Although the hydrological community recognizes many sources of uncertainty, their magnitude and impact on the final streamflow prediction are usually not evaluated. It is worth noting that prediction systems issuing a single-valued forecasts are called deterministic, because they lack evaluating uncertainty even when the single prediction originates from an ensemble of forecasts.

In the case of white or gray-box models, predictions usually emanate from a single propagation of the initial state of the catchment that represents the "best" prediction of precipitation and a priori status of evapotranspiration, soil moisture and other system variables according to the judgement of the modeller and the information at hand. In the case of black-box models, the modeller defines the complete system (data, inputs and model) as the "best" representation of the hydrological dynamics of the basin.

I.3 Hydrological ensemble prediction systems

Over the last ten years, the paradigm of the Ensemble Prediction System (EPS) emerged in the hydrological community, notably with the start-up of an international project called "The Hydrologic Ensemble Prediction EXperiment (HEPEX)" [130] and momentum generated by the EPS development in atmospheric and meteorological communities.

This "new" paradigm is based on the basic principle: "two heads are better than one" i.e. that bad choices can be avoided by encouraging different viewpoints to reach mutually agreeable decisions [36].

Additionally, actual computational resources offer the possibility to build HEPS, propagating multiple plausible initial conditions. Figure I.1, based on scheme presented by Moffet et al. [107], schematically shows different sources of uncertainty: the left oval represents an envelope of all the possible initializations of the system, dark circles within it illustrate some possible initialization of the system that are assumed to capture the aspects of the true initial Probability Density Function (PDF) taking into account the uncertainty of the variables within the system and the assimilation of the information provided. Holding this assumption as true, we can obtain a HEPS (stars in the right set) projecting the initial states by running one or multiple white, gray, or black-box models with one or more datasets and model parameters simulating the structural and parametric uncertainty of the hydrological process. Each **member** from the forecast PDF is then considered a member of the **ensemble** of forecasts.

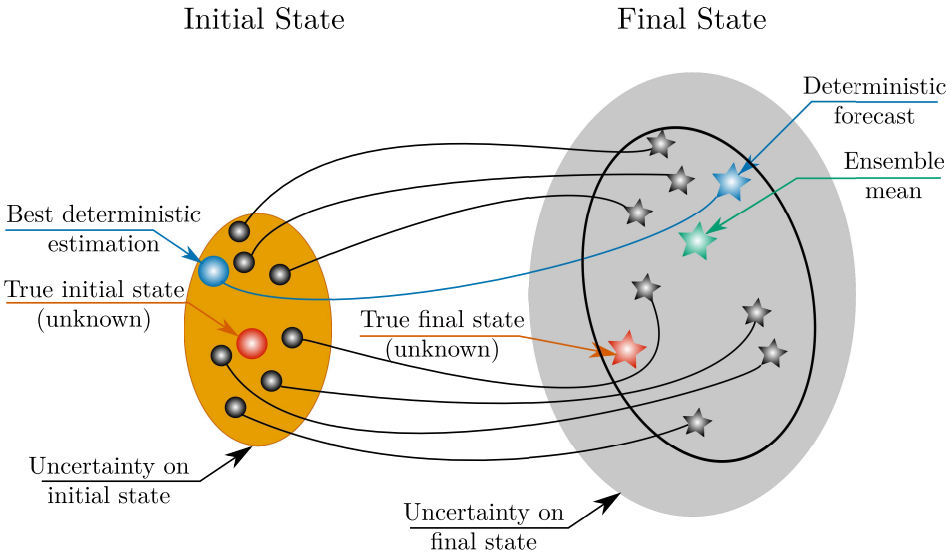


Figure I.1: Some concepts about HEPS.

The more plausible system initializations are picked, the greater the probability of resolving the initial state of the system. Plausible system initializations must be perceive as scenarios consistent with observations and the unknown true state of the system. This is why there is

a lower confidence in forecasts derived from a deterministic set-up, which technically has an initial and forecast PDF comprised of one member losing all uncertainty information of the process.

Importantly, on average situations, it is generally accepted that the mean prediction outperforms the best individual predictor and the deterministic model [7, 148, 158]. One downfall to the ensemble mean is that if the forecast PDF supports two or more statistical modes[†], the mean will combine these solutions and remove this information from the forecaster. However, beyond the mean prediction, the participation of several experts or models in a decision context raises questions such as:

- How different are the viewpoints of the models (a property named resolution)?
- What accuracy may be attributed to a multimodel system (an attribute called bias)?
- What is the agreement between the individual model responses and the observed event (a feature known as system reliability)?

In a hydrological context, answers to these questions and others not less important lead us to promote HEPS as a valuable solution. For example, a hydrologist could always undermine the ensemble approach by resorting to "expert knowledge" for identifying the best possible scenario. However, this decision would be questioned in relation to the rapidly increasing computational capabilities and the lack of uncertainty evaluation. Note that probabilistic forecasting or ensemble prediction hold no intrinsic value. They acquire value through their ability to influence decisions [111].

In summary, in HEPS we seek to account for the effect of errors and/or uncertainty in the different hydrological sub-processes. Thus, in the absence of an explicit form of account, such uncertainty and therefore the unknown PDF in the initial conditions are verified. A simple method consists in generating and propagating initialized plausible states of the system, which leads us to the so-called Uncertainty Cascade Model, proposed by Pappenberger et al. [115], which identifies different sources of uncertainty as a combinatorial problem (Fig. I.2) or analogously in the black-box context to the Multi-Level Diversity (MLD) model (Fig. I.3) described below. So, we **assume** that our synthetically derived initial PDF **captures** the aspects of the true initial PDF and consequently the forecast PDF has a strong likelihood of capturing what is verified.

I.4 Multi-level diversity model

The mathematical concept of **diversity** favours the complementarity of the members of an ensemble: if a member has a low performance according to some criteria and some strength in another dimension of the problem, there exist one or more members that can minimize the

[†]The mode of a continuous probability distribution is the value x at which its probability density function has its maximum value, so, informally speaking, the mode is at the peak.

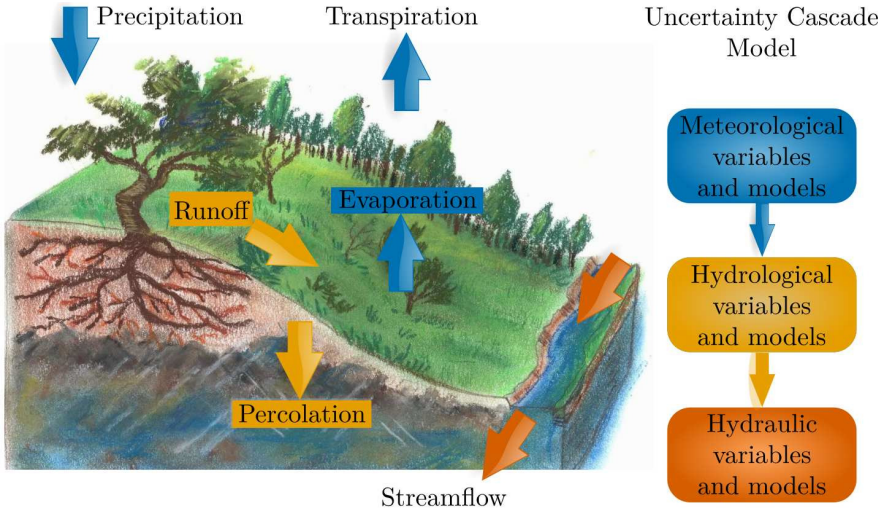


Figure I.2: Hydrological scheme and Uncertainty Cascade Model.

flaw of such member, ensuring that the final result is improved, i.e. resorting to many members of different strengths to support mutually agreeable decisions, moving beyond the desire to find a single model exempt from errors.

The diversity concept has been studied explicitly in the community of machine learning, more precisely in the Multiple Classifier System (MCS) approach [92]. Indeed, this community has established a MLD model to promote the construction of ensembles (Fig. I.3). This model forces the diversity in ensembles based systems combining the use of different data subsets, models with different input subsets, different models and/or different parameter settings, even including a combiner level in order to optimize the final ensemble based on members selection or fusion [92].

Diversity can be efficiently used for building ensembles in an intuitive or implicit manner, as unequivocally demonstrated by the AdaBoost algorithm [136, 139]. Although various measures have been proposed to explicitly quantify this factor in the process of building the ensemble, it does not share the success of more implicit methodologies [92]. The importance of diversity may seem obvious but its relation with ensemble prediction properties is not.

I.5 Importance of HEPS

The importance of HEPS can be evaluated from conceptual and operational perspectives. HEPS conceptually represents the possible integration of different views, becoming the ideal scheme for the combination of different types of modelling that are often seen in opposition.

HEPSs, as a "new" paradigm in hydrology, stand aside from the paradigm of perfect model exempt of any error, accepting that uncertainty makes active part of the prediction process

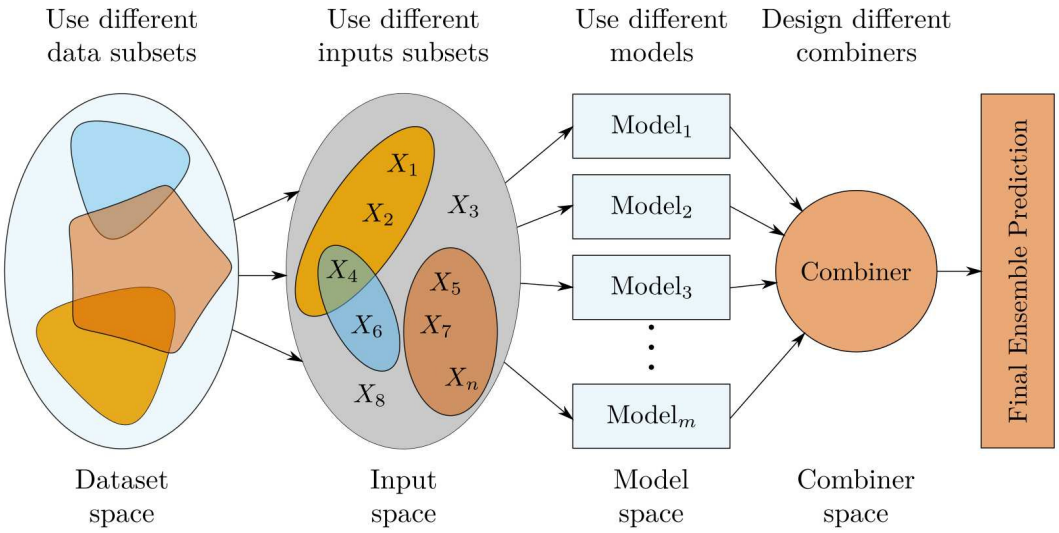


Figure I.3: Multi-Level Diversity model.

and should therefore be reflected in the final prediction to assess the risk in decision making. In this context, the operational advantage of HEPS is obvious since the availability of multiple scenarios, which are **supposed** to accurately represent hydrologic variability phenomenon, is the basis for improved operational water management and a better anticipation of hydrologic extremes. Such forecasting and warning systems have been developed and applied to improve flood control and drought risk planning, as well as to optimize water management and regulation for different economic uses [124].

Figure I.4 shows an example of a HEPS associated to a Decision Support System with different warning levels. Forecasts (members) are issued from day 19, for the next nine days or nine FTH. On day 24, there are already many scenarios reaching the low warning level, which should enable certain actions by the appropriate authorities. On day 25, two scenarios extend to the high warning level and from day 26 to 28 some scenarios even reach the extreme level warning. The PDF of the twenty-member forecast for day 29 is drawn on the left side on the figure. Now, if one considers the following Flood Contingency Plan (FCP) example for a decision-making process: the cost of not activating a FCP since it does reach the High Warning Level (HWL) is \$1M, while the cost of activating the FCP is \$0.05M. If the probability of a streamflow higher than 285m^3/s (HWL) is 10% (2 scenarios out of 20), *So what would a decision maker should do?*

Based on the reliability of the HEPS, i.e. taking all cases in which the event is predicted to occur with a probability of x%, that event should occur exactly in x% of these cases; not more and not less, the decision maker can reach the following preliminary analysis based on the evaluation of 100 cases where the scenario occurs: Cost for no activation of the FCP = \$10M ($10 \times$ \$1M). Full activation of FCP = \$5M ($100 \times$ \$0.05M).

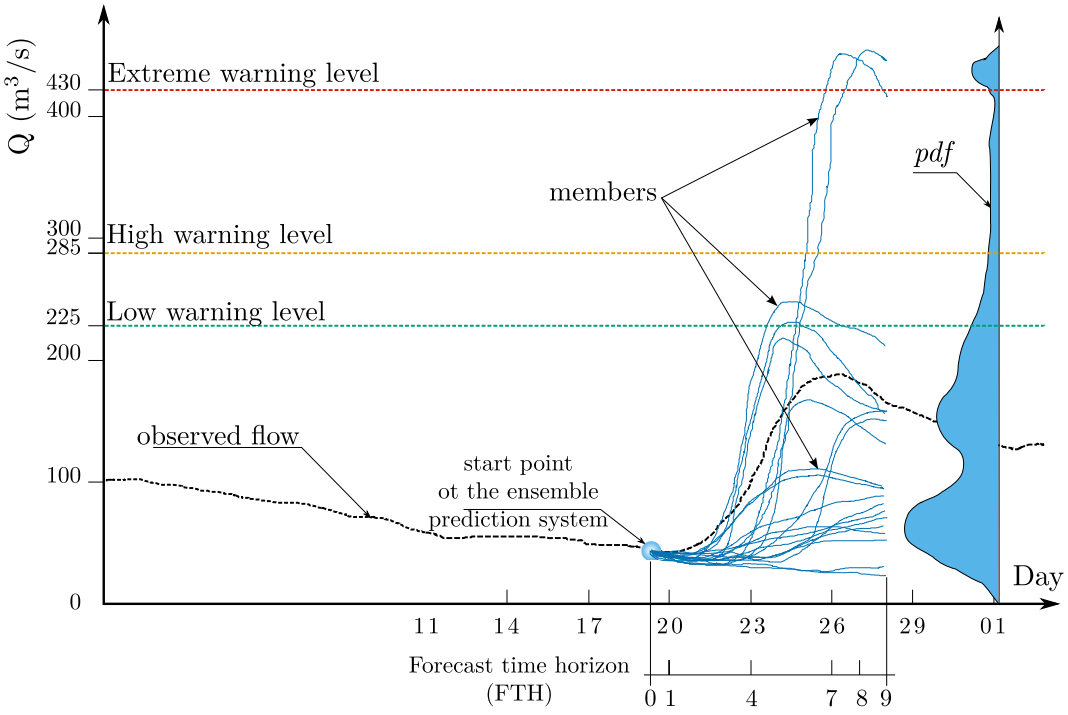


Figure I.4: HEPS as basis of a Decision Support System.

So, we can conclude that it is advantageous to activate FCP if the probability of a HWL is higher than 5%. The cost-loss ratio is directly proportional to the probabilities needed in order to benefit from acting on the forecast. Clearly, part of the responsibility of a correct decision relies on the reliability of the HEPS. In brief, a probabilistic prediction encourages a reaction when the probability of a specific event exceeds a threshold defined by the end-user. However, the HEPS community has highlighted that despite the proven high HEPS performance, adoption and evaluation of such information by a committee of alarms is a highly complex task given the new prediction format, the assimilation of new tools to be applied and how the reliability is transmitted to the end-users [124, 158].

I.6 Thesis main topics

Based on the diversity concept, we investigate two directions for HEPS development. The first one explores a HEPS conceived with the partial application of the Uncertainty Cascade Model with the combiner level exploited in the MLD model, while the second focuses on building ANN ensembles with implicit diversity partially using the MLD model.

I.6.1 The complexity of the HEPS as an operational barrier

One can easily visualize that HEPS can be generated from the combination of scenarios from different sources of uncertainty regardless of the type of modelling. In the case of physical and

conceptual models, the most common way is the use of numerical weather prediction forecasts, which are then used as input to hydrological simulation models [45]. Alternatives include additional model parameter uncertainty [115] and multi-hydrological models approaches [148]. Also, other mechanisms increase the range of possibilities such as: pre-processing techniques [64, 131], weather forecasting model resolutions [102], radar blending [116] and data assimilation techniques [95].

In the case of black-box models, the MLD model (Fig. I.3) shows that the ability to launch a HEPS with thousands of members is easily achievable, simply combining multiple subsets of data, different configurations of the input space, a wide range of mathematical models, and variations in the configuration of each of them.

But the combination of scenarios, independent of the type of model, is limited by the computational capacity of forecast centres. Hypothetically, individual or joint evaluation of each of the sources of uncertainty can put the system to the limit of available computational resources.

In this regard, Cloke and Pappenberger [45] have highlighted the high computational demand of coupling a Meteorological Ensemble Prediction System (MEPS) to a hydrological model. But, He et al. [78] and Bao et al. [17] have shown that the combination of the information derived from many MEPS improve early flood warning systems. Moreover, if the parametric uncertainty of hydrological models is assessed under the principle of equifinality [19] and if the structural uncertainty is tackled through a multi-model approach, the number of scenarios in the uncertainty cascade model may rapidly turn out to be quite large. Simplification of such a HEPS inevitably becomes a mandatory step from an operational standpoint.

Consider the complexity related to the number of members of the two hypothetical examples presented in Table I.1. Conformation of both HEPS is based on the individual uncertainty components that have been evaluated by different authors. Also, under the cooperative philosophy that we want to promote in this thesis, the combination of both types of modelling can lead to worsen the manageability of such systems.

In summary, the examples given in Table I.1 are intended to show the present and future need of resorting to simplification methods for assessing streamflow uncertainty without sacrificing the quality of probabilistic predictions.

I.6.2 ANN acceptance and the HEPS opportunity

In hydrology, the efficiency of ANN as regression model has been demonstrated in numerous studies, syntheses of these advances can be found in Abrahart et al. [5], Maier and Dandy [97], Maier et al. [98], and Abrahart et al. [6]. With regard to the acceptance of the ANN in an operational hydrological context, Abrahart et al. [5] presented the so-called "Revised Technological Adoption Life Cycle (RTALC)" (Fig. I.5). This model describes the market

Table I.1: Examples of hypothetical HEPS complexity.

HEPS	Uncertainty component	Description	No scenarios	References
Based on gray-box models	Meteorological	Precipitation derived from ten global meteorological centres	259	[23]
	Hydrological conceptualization	Multiple gray-box models	16	[148]
	Model parameters	Parameter sets due to the GLUE methodology	6	[115]
	Final number of members		**36864**	
Based on black-box models	Datasets	Using bagging, boosting (i.e. stratified sampling of the original training set), or learning vectors identification	500	[8, 43]
	Model structure	ANN structures such as: Feed-Forward, Elman, Fully recurrent and Echo state networks	9	[149]
	Model parameters	Random initializations for training	50	[10]
	Final number of members		**225000**	

penetration of a new technological product in terms of progression with respect to the type of consumers that it attracts throughout its useful life. It identifies five stages separated by gaps that proportionately represent the difficulty of moving to the next stage. So, it presents the smooth transition between innovators who are synchronized with the latest technological developments, and the Early Adopters who are not technologists but nonetheless find it easy to understand and appreciate the potential rewards. In contrast, the transition between the latter and the Early Majority or Pragmatists stands out as the most difficult phase since it must be shown that the product is not part of a fad and that it is highly practical. It is precisely at this stage that Hydroinformatics community comes focusing its efforts to gain the momentum that would open the way to the Late Majority, who will wait until a particular technological development has become an established standard with lots of support. The RTALC model finally illustrates the weak transition to the Laggards who, for some reason or other, simply don't want anything to do with the latest technological innovations.

The worldwide trend of global acceptance of the HEPS[‡], who implicitly accept the need of additional and complementary formulations in hydrological forecasting, is seen as an ideal opportunity for the ANN to close the gap between Early Adopters and Early Majority.

The greatest opportunity to launch ANN in operational hydrology consists in the development

[‡] http://hepex.irstea.fr/operational-heps-systems-around-the-globe/.

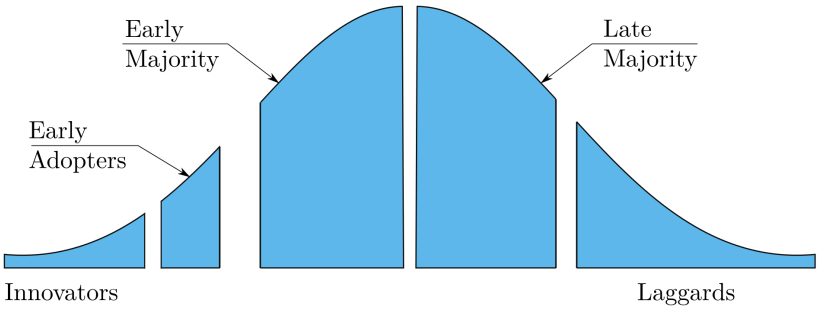


Figure I.5: Revised Technological Adoption Life Cycle (RTALC).

of hybrid solutions (ANN-hydrological models) that would allow greater visibility within the community through the combination of the uncertainty cascade model (Fig. I.2) and of the MLD model (Fig. I.3). We adopt in this thesis the development of ensembles based solely on ANN to evaluate the MLD model, prioritizing methodological simplicity for highlighting diversity as a fundamental concept.

It is also important to note that in the ensemble modelling context, similarities in theoretical developments in machine learning and hydrometeorology abound and are deemed complementary. Machine learning groups many theoretical concepts underlying the benefits of multimodel schemes, notably based on the behaviour of bias, variance, covariance, and diversity between ensemble members [90]. At the same time, the hydrometeorological community has developed probabilistic metrics, called **scores**, used not only to evaluate the "most likely" simulation or prediction but also their uncertainty.

In the general case of the black-box models, it is clear that the ensemble trend has not yet had much impact on the hydroinformatics community, where the main emphasis is focused on the improvement of individual models or the adoption of increasingly sophisticated techniques in simulation. However, machine learning community highlights the MCS approach, which under the MLD model promote the construction of ensembles to address typical problems in pattern recognition [92]. One should thus expect a transfer of knowledge to the hydroinformatics community to give greater popularity to HEPS based on this philosophy.

Some ANN studies already promote the evaluation of uncertainty in simulation or alternatively accept the ensemble approach as a cooperative mechanism for reducing the simulation error [8]. In this aspect, Boucher et al. [22], from the analysis of the evolution of the ANN training process, demonstrated that although the ensemble modelling reduces system bias, the reliability may be severely compromised. Nix and Weigend [114] presented an evaluation of the simulation uncertainty from the estimation of the mean and the variance of the target as a function of the input, given an assumed target error-distribution model. Likewise, several authors presented different methods for evaluating the ANN output uncertainty from the construction of prediction and confidence intervals [44, 82, 84, 85, 136].

These studies have a common denominator: the active search of an "adequate" variability or diversity, which is coherent with the philosophy of the MLD model. Consequently, in the second part of this thesis, we prioritize the use of simple models and techniques in the construction of a HEPS based on ANN, seeking transparency in the implementation of the MLD model and an easier adoption of our proposal in the hydrological community.

I.7 Hypothesis and objectives

I.7.1 HEPS simplification

In the HEPS, we seek to capture the uncertainty associated with the prediction. For this purpose, it is not useful to have a 200 000-member ensemble if all the members lead to an identical solution. At the same time, it is not ideal to have a 10-member ensemble with solutions exhibiting no correlation. Ideally the output will produce significant difference in solutions whose forecast distribution matches the actual frequency of occurrence.

To evaluate the number of members required in a specific HEPS, the first part of this thesis explores different simplification schemes applied on a complex HEPS designed by Velázquez et al. [148], who showed the relevance of combining two sources of uncertainty in hydrological forecasting: sixteen lumped hydrological models driven by the fifty weather ensemble forecasts from the European Centre for Medium-range Weather Forecasts (ECMWF), resulting in an 800-member HEPS. But they also highlighted that such HEPS complexity may become an operational burden when one has to evaluate several hundreds of scenarios at each time-step.

We thus propose searching optimal selection of predictors according to the probabilistic behaviour of the system. This problem is easily associated to the response combiner level of the MLD model (Fig. I.3). In this case, the simplification or selection process is known as "overproduce and select", where the hypothesis of improvement of the HEPS is based on the existence of an optimal combination of predictors that minimizes the ensemble mean error with respect to the observation. This is a property known as the bias of the system. It maximizes system reliability, property that is a function of the predictors dispersion.

Based on Fig. I.1, that showed the concepts in the HEPS production, Fig. I.6 schematically illustrates the hypothesis about optimal selection of predictors or members. In this case the selection of the three members represented by stars enveloped with circles shows that the optimized average value is closer to the true final state that the average value of the original ensemble. However the problem is not confined to the vicinity of the average value to the observed value (bias) because the dispersion of the members is a complex function over other probabilistic properties of the system such as reliability, resolution and consistency, between others. Note that selecting a member subset not only decreases complexity but can also improves ensemble quality.

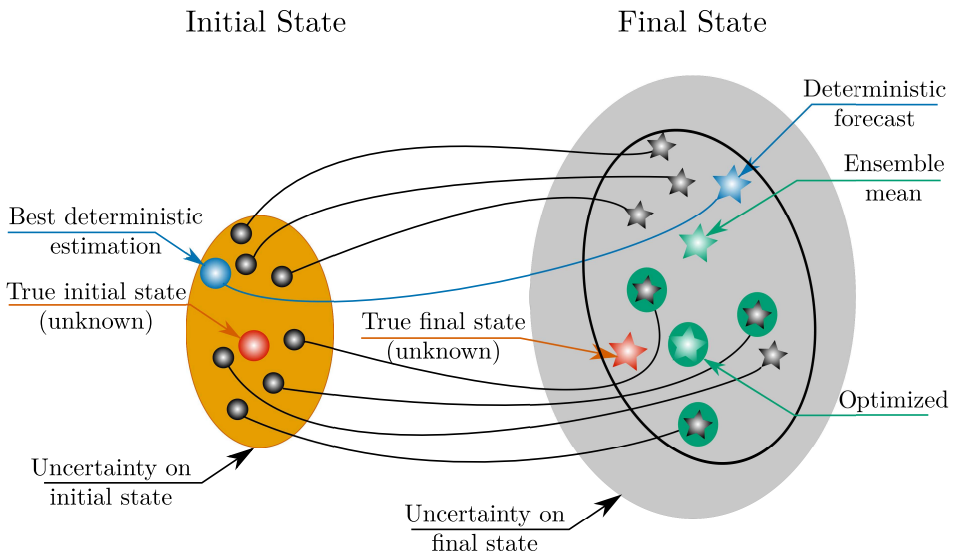


Figure I.6: Members selection scheme.

The member selection problem, considered explicitly at the combiner level in the MLD model, leads to one of the most relevant topics in machine learning community: **feature selection**. That is, we are interested in finding k of the d members that give us the most information and we discard the other $(d-k)$ members, consequently the possible approaches for members or predictor selection are the same. In this context, we elucidate the following **objectives in the predictor selection process**:

- To evaluate the tradeoff between several ensemble probabilistic properties related to the mean ensemble value and the predictors distribution.
- To determine the optimal number of members in the HEPS in function of its complexity and performance.
- To estimate the ability of extrapolating the simplification scheme in several FTHs and another basins.
- To define the advantages and the weakness of different selection tools developed in the machine learning community.

I.7.2 Building ANN ensembles as HEPS

The few examples of ANN ensemble applications for hydrological forecasting usually focus on minimizing the error of the ensemble mean response (bias) penalizing a high variance. However, our hypothesis on the construction of a HEPS with ANN is mainly based on finding an "adequate" variability on two premises: the active imposition of ensemble diversity using partially the MLD model and the use of simple FFNN that has the advantages of ensemble modelling in contrast to the use of more complex ANN structures. It is worth noting that in the prioritization of a minimum level of complexity we adopt an ensemble of FFNNs, neglecting

structural variability due to the choice of a particular ANN structure. Also we ignore the combiner or post-processing level of the MLD model where one can focus the optimization of the HEPS, as presented in the previous section.

Regarding HEPS based on ANN, the closest work in our line of "adequate" response variability of the ensemble is the one performed by Boucher et al. [21], who emphasized that an ensemble for which all individual members have the same configuration (train datasets, inputs, and structure) leads to a lower reliability in the prediction ensemble and under-dispersed results. In an attempt to improve the diversity of the ensemble, they argued that the lack of diversity was due to the final stages of the ANN training algorithm. So, they proposed to seek ensemble diversity integrating networks optimized at each training iteration or epoch [22].

Here, we propose an evaluation of an ANN ensemble or stack, where variability is the result of training each independent ANN or member with different stratified sub-samples, different input system schema, and finally, different parametrization of each ANN since the initial conditions are random and a local search algorithm is used for ANN training. The skill of the proposed system is confronted to a the baseline model consisting of a Ensemble of 30 FFNNs trained with early stopping using a **R**andom sampling of **100 P**ercent of the available information and a single predefined set of inputs variables (R100P).

As it will be show later in Chapters 6 and 7, although the baseline model does not represent the confluence of the latest advances in some ANN topics as data selection, input variable selection, training algorithms and ANN structures, the use of the **ensemble approach** with simple FFNNs is demonstrated as efficient as any other individual structures, much more sophistical evaluated by Vos [149] on the same basins chosen here, which are part of the MOdel Parameter Estimation eXperiment (MOPEX) project; databases that are freely distributed[§].

Another very important feature of the evaluation of our hypothesis is the coverage of topics considered important in ANN in hydrology literature [1, 6, 97, 98] such as: the establishment of clearer protocols of experimentation, the use of benchmark datasets that enable verification, the tests for bias related to different data partitioning, a multicriteria framework, and finally an uncertainty analysis.

Consequently, in this thesis the development of the ANN ensembles, in relation to its deterministic and probabilistic performance, is based on the following **objectives**:

- To estimate the gain of promoting different diversity sources established in the MLD model.
- To quantify the impact of the length of the data-series used in the ANN training, above all in terms of their informativeness.
- To calculate the effect of resampling methods for configuring the datasets required for an early stopping method of ANN training.

[§] `http://www.nws.noaa.gov/oh/mopex/mo_datasets.htm`

- To evaluate an Input Variable Selection schema encouraging diversity by adding variables one by one (stepwise selection) in different sub-samples evaluations.

I.8 Thesis structure

This thesis is divided into four parts: the first addresses the basic concepts in Chap. 1 and 2. Chapter 1 reviews the basic concepts about multi-model prediction, including measures such as bias, variance, covariance, and diversity, plus pertinent scores developed in the hydrometeorological community to evaluate the quality of the prediction ensembles. Chapter 2 briefly documents some concepts and tools developed in the machine learning community. The generalization concept is presented as a key aspect of modelling. The datasets nomenclature that is used in the experimental design of this thesis is also detailed, as well as an overview of clustering, regression, and features selection problems.

The second part, subdivided in Chap. 3, 4, and 5, presents simplification schemes of a HEPS of 800 scenarios or hydrological members resulting from the combination of sixteen hydrological models and fifty rainfall forecasts from the ECMWF-EPS, which corresponds partly to Uncertainty Cascade Model (Fig. I.2).

Chapter 3 assesses the degree of simplification, i.e. the reduction of the number of hydrological members that can be achieved in terms of complexity and performance. Here, a stepwise method and a combined criterion are proposed to evaluate the simplification scheme at the 9th FTH. So, we find a subset that offers similar or better performance than the reference set of 800 hydrological members. The subset of hydrological members serves to define the importance of each hydrological model or the Hydrological Models Participation (HMP).

Chapter 4 explores the efficiency of the simplification schemes. Their generalization ability is confronted with other FTHs and neighbouring basins. Tests are made in two ways. At the local level, the transferability of the selection scheme assessed 9-days ahead is evaluated for the other eight FTH. At the regional or cluster level, the analysis evaluate a new simplification scheme based on a proposal from a regional integration mechanism, tested in neighbouring basins.

Chapter 5 compares the performance of various optimization schemes. Given the 9-day lead time for a catchment, the HMP is sought from four techniques: Linear Correlation Elimination, Mutual Information, Backward Greedy Selection (BGS), and Nondominated Sorting Genetic Algorithm II (NSGA-II). The HMP will specify the number of representative members to propagate into each hydrological model, while generalization is evaluated in a neighbouring catchment at different forecast time horizons.

In the third part, Chap. 6 and 7, we evaluate partially the MLD model (Fig. I.3) with a stack or ensemble of 30 ANNs. Chapter 6 presents a methodology to evaluate stratified sub-

samples as representative datasets. Subsequently an estimation of the impact of the length of observed records is presented. Chapter 7 proposes a framework based on two separate but complementary topics in ANN development: data stratification and Input Variable Selection (IVS). Each predictor is trained based on input spaces defined by the IVS application on different stratified sub-samples. All this, added to the favourable variability of classical FFNN optimization, leads us to our ultimate goal: diversity in the prediction.

Finally, the fourth part gathers the general conclusion and contributions of the thesis and proposes some guidelines for future work.

Part I

Basic Concepts

Chapter 1

Ensemble Prediction System Evaluation

The evaluation of EPS, i.e. systems with a pool of predictors instead of only one, can be accomplished in two ways: deterministically or probabilistically. Deterministic evaluation, in its simplest version, considers the average of the predictions as the system output, while a probabilistic (distribution-oriented) approach is based on the notion that the joint distribution of forecasts and observations contains all of the non-time-dependent information relevant for evaluating forecast quality [111].

In this chapter, we first introduce the EPS manipulation from a probabilistic perspective. We begin by reviewing some concepts derived from the basics of models based on a single predictor. Such concepts are extended to different scenarios accounting for the variability of calibration data and/or model structures. Thus, we show one of the fundamental aspects of multi-model approach: the possibility of reducing the error by manipulating the ensemble covariance.

However, from a probabilistic viewpoint, square error reduction is not a sufficient criterion for increasing the prediction quality, it is also imperative to obtain ensemble PDFs that exhibit appropriate coverage and are as sharp as possible [66, 155]. Consequently, we discuss other important characteristics such as reliability, resolution, sharpness, and consistency. Finally, we describe several mathematical tools designed to evaluate these properties.

1.1 Bias, variance, and covariance

In order to lead the reader to appreciate three basic concepts used in this work, that is: bias, variance, and covariance, we present the interpretation of Mean Square Error (MSE) in various scenarios.

Traditionally, forecast verification has consisted in the computation of measures of the overall (average) correspondence between forecasts and observations (e.g. through MSE). This traditional measure-oriented approach tends to focus on one or two overall aspects of forecast quality, such as accuracy and skill, if another system serves as reference [111].

1.1.1 Single-predictor model

In the simplest context of deterministic prediction, i.e. including only one predictor, the analysis focuses on the comparison of two scalar values at each time-step: one observation versus one prediction. In this case theMSE is a function of the sample variance of the forecasts and observations, the bias or mean difference between the observed and forecasted values, and the sample covariance between the observations and forecasts [112] (we explain the full details of their proof in Appendix B):

$$\begin{aligned} \mathrm{MSE}(\mathbf{o},\mathbf{y}) \;=\; & \underbrace{\frac{1}{N}\sum_{t=1}^{N}\left(y^t-\bar{\mathbf{y}}\right)^2}_{\text{forecasts variance}} + \underbrace{\frac{1}{N}\sum_{t=1}^{N}\left(o^t-\bar{\mathbf{o}}\right)^2}_{\text{observations variance}} + \underbrace{(\bar{\mathbf{y}}-\bar{\mathbf{o}})^2}_{\text{bias}} \\ & \underbrace{-\,2\,\frac{1}{N}\sum_{t=1}^{N}\left(y^t-\bar{\mathbf{y}}\right)\left(o^t-\bar{\mathbf{o}}\right),}_{\text{covariance}} \end{aligned} \tag{1.1}$$

where superscript t represents each of the N observations and forecasts (o^t and y^t). In this thesis we adopted bold lowercase and uppercase letters to represent vectors and matrices respectively. Additionally an over-line symbolizes the mean value. In this case all of the N observations and forecasts are indicated by $\mathbf{o}$ and $\mathbf{y}$.

The complex relationship between these factors can be addressed from a multicriteria optimization, as proposed by Gupta et al. [70]. They used an alternative decomposition to show that, in order to minimize MSE, the variability has to be underestimated. Additionally, they evaluate the relationship between these components and the overall volume of flow, the spread of flows and the timing and shape of the hydrograph.

1.1.2 Bias-variance dilemma

If we have a model y for some data x that may provide a very good fit for a specific sample χ, but not for another one (a model with a poor generalization where for new data the prediction and the "true" observed value are very different), bias measures the accuracy or the quality of the match: high bias implies poor match. Another way of measuring the "match" is the variance in the prediction errors, that is the precision or specificity of the match: a high variance implies a weak match between o and $y(x)$. One can adjust the bias and variance of its model but the bias-variance relation states that the two terms are not independent. To

quantify how pertinent a model $y(\cdot)$ is, we must average over many possible datasets [7]. So, the expected squared error can be written as follows:

$$E_\chi\left[\left(E\left[o|x\right]-y\left(x\right)\right)^2|x\right]=\underbrace{E_\chi\left[\left(E\left[o|x\right]-E\left[y(x)\right]\right)^2|x\right]}_{\text{bias}^2}+\underbrace{E_\chi\left[\left(y(x)-E\left[y(x)\right]\right)^2|x\right]}_{\text{variance}}. \quad (1.2)$$

In Appendix C we present the proof of Eq. 1.2, proposed by Geman et al. [65], which is known as the *bias/variance dilemma* or generalization error components. Generalization is a common issue for all types of data-driven models and aims at reducing modelling errors [8]. The *bias* represents how close, on average, our model responds to the observed "true" value. The variance indicates how 'stable' the model response is given slightly different calibration data. Bias and variance usually work in opposition to each other: attempts to reduce the bias component will often cause an increase in variance, and vice versa. [7, 36].

Given the importance of this dilemma, consider the trivialization of these concepts in the following example proposed by Moore et al. [110]. In Fig. 1.1 our model is a dart shooter and the objective is to score a bull's-eye. Bias shows how far the average shoot is from the bull's-eye. Variance depicts the dispersion of the darts – independent of their position on the dartboard. The ideal situation clearly consists in low bias and low variance.

Figure 1.1: The dartboard and the bias-variance analogy.

It is worth noting that, in this context, the bias and variance, although statistically representing the same concept of Eq. 1.1, are analyzed in a different fashion, while Eq. 1.2 was only developed in the "longitudinal" domain of the series (the time domain), the bias-variance dilemma takes into account also a "transverse" dimension corresponding to the variability of both data and model parameters.

1.1.3 Multi-predictor model

We can take into account the possibility that our model could be an ensemble of D individual models, $\mathbf{Y}=\{y_1,y_2,\ldots,y_D\}$, using a linear combiner. In this case, Krogh and Vedelsby [90] proposed the *ambiguity decomposition* or *accuracy-diversity breakdown*, which establishes that

the MSE between mean forecasts $\bar{\mathbf{y}}$ and observations $\mathbf{o}$ can be broken into two components:

$$\mathrm{MSE}(\bar{\mathbf{y}}, \mathbf{o}) = \underbrace{\frac{1}{N}\sum_{t=1}^{N}\left(\frac{1}{D}\sum_{d=1}^{D}\left(y_d(\mathbf{x}^t) - o^t\right)^2\right)}_{\text{individual model errors}} - \underbrace{\frac{1}{N}\sum_{t=1}^{N}\left(\frac{1}{D}\sum_{d=1}^{D}\left(y_d(\mathbf{x}^t) - \bar{y}(\mathbf{x}^t)\right)^2\right)}_{\text{predictors interactions}}, \tag{1.3}$$

where $\mathbf{x}^t$ represents the model inputs at time t. Herein the t superscript represents the time-step, the d subscript designates one of the D models evaluated in the prediction ensemble, and $\bar{y}(\mathbf{x}^t)$ represents the ensemble mean forecast at time t ($\bar{y}(\mathbf{x}^t) = \frac{1}{D}\sum_{d=1}^{D} y_d(\mathbf{x}^t)$). The first term on the right hand side of Eq. 1.3 is the average squared error of the individual models, while the second term quantifies the interactions between the predictions. Note that this second term, the "ambiguity", is always positive. This guarantees that, for an arbitrary data point, the ensemble squared error is always less than or equal to the average of the individual squared errors [38]. The larger the ambiguity term is, the larger the ensemble error reduction. However, as the variability of the individuals rises, so does the value of the first term. This therefore reveals that diversity itself is not enough, we need to get the right balance between diversity (the ambiguity term) and individual accuracy (the average error term), in order to achieve the lowest overall ensemble error [36].

1.1.4 Bias-variance-covariance dilemma

Now, consider that the d^{th} model of our multi-predictor model is calibrated using a specific dataset. Taking the expected value of Eq. 1.3 over Z training sets, we obtain the Bias-Variance-Covariance decomposition proposed by Ueda and Nakano [146]. This decomposition expresses that the expected squared error of an ensemble from a target o is given by:

$$\begin{aligned} E\left[(\bar{\mathbf{y}} - o)^2\right] &= \overline{\mathrm{bias}}^2 + \frac{1}{Z}\overline{\mathrm{var}} + \left(1 - \frac{1}{Z}\right)\overline{covar}, \\ \text{where:} \quad & \overline{\mathrm{bias}} = \frac{1}{Z}\sum_i \left(E_i\left[y_i\right] - o\right), \\ & \overline{\mathrm{var}} = \frac{1}{Z}\sum_i E\left[\left(y_i - E_i\left[y_i\right]\right)^2\right], \\ & \overline{\mathrm{covar}} = \frac{1}{Z(Z-1)}\sum_i\sum_{j\neq i} E_{i,j}\left[\left(y_i - E_i\left[y_j\right]\right)\left(y_j - E_j\left[y_j\right]\right)\right]. \end{aligned} \tag{1.4}$$

Full proof can be found in [36]. While the bias and variance terms are constrained to be positive, the covariance between models may become negative. We can see that the error of an ensemble of models depends critically on the level of error correlation between them, quantified in the covariance term. We would ideally like to decrease the covariance without causing any increase in the bias or variance terms [36] – the definition of diversity thus emerges as an extra degree of freedom in the bias-variance dilemma. This extra degree of freedom allows

an ensemble to approximate functions that are difficult (if not impossible) to find using a single model [38].

So far, we have revealed the advantages of a prediction scheme that takes into account the variability (diversity) coming from several fronts: the calibration datasets, the sets of parameters that govern a particular model, and finally the multimodel approach. However, the previous approaches have only explicitly highlighted three statistical properties of ensemble prediction systems: bias, variance, and covariance.

1.2 Ensemble forecasts quality

Murphy [111] presented three distinct types of goodness in ensemble forecasting:

1. The correspondence between the forecasters' judgements and their forecasts (i.e. forecasts consistency*);
2. The correspondence between the forecasts and the matching observations (i.e. quality); and
3. The economic and/or other benefits through the use of the forecast (i.e. value).

Even if forecasts consistency and value justify the use of an EPS, it is clear that ensuring the EPS quality is the first step towards their overall goodness evaluation.

The distribution-oriented approach reveals that forecast quality is inherently multifaceted in nature [45, 111]. In the following, we quote some of the properties commonly evaluated in probabilistic forecasting. The reader is referred to Wilks [161] for a more detailed description of these features. Note that we use Gaussian distributions in the following examples for the sole purpose of facilitating the explanation.

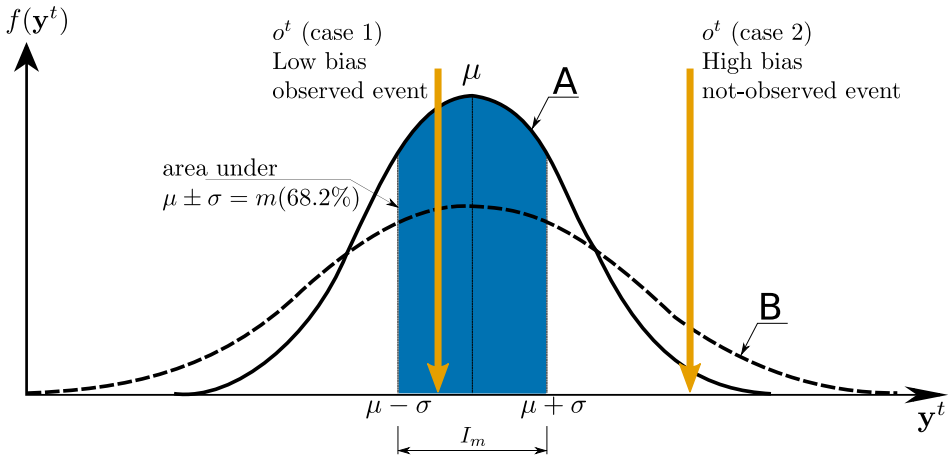


Figure 1.2: Probabilistic forecasting evaluation at time t.

*Here we speak of forecasts consistency to avoid confusion with the ensemble property called consistency.

- Bias: also called unconditional bias or systematic bias, measures the correspondence between the mean forecast and the mean observation. In Fig. 1.2 we illustrate two cases regarding PDF *A*. The observation o^t first occurs near the mean forecast[†] (case 1), i.e. a system with low bias. Second, the observation o^t is located further from the central value (case 2), i.e. a highly biased system.
- Reliability: relates to the occurrence of event o^t given a probability threshold m, averaged over all N observation-forecast pairs.
  Consider in Fig. 1.2 the reliability evaluation of a threshold probability equal to $m = 68.2\%$ in PDF *A*. At each time-step, it must be determined whether the event falls or not into the I_m region bounded by this probability value. Subsequently, the conditional observed frequency $\bar{o}_m$ is evaluated for the N observation-forecast pairs:

$$\bar{o}_m = \frac{1}{N}\sum_{t=1}^{N} r^t \quad \text{where} \quad r^t = \begin{cases} 1 & \text{if } o^t \in I_m \\ 0 & \text{otherwise} \end{cases} \tag{1.5}$$

  For this probability threshold, the system is perfectly reliable if $\bar{o}_m$ is equal to m (imagine that in 1000 cases, the event fell 682 times within the intervals evaluated for a probability of 68.2%). We say that the system is overforecasting if $\bar{o}_m$ is less than m and underforecasting otherwise.
  Given that m denotes the different M thresholds of probability to assess, the reliability of the system can be directly measured from the comparison of these thresholds with the M observed conditional probabilities. The goal is to have well-calibrated forecast systems for which the relative frequency is essentially equal to the probability of the forecast, i.e. $\bar{o}_m = m$ (See Section 1.3.3 for more details).
- Resolution: consists in the difference between this same conditional mean observation ($\bar{o}_m$) and the overall unconditional mean observation ($\bar{o}$); again, averaged over all forecasts. An important question to consider is then: to what extent do the conditional means of the observations corresponding to the streamflow forecasts of $3\text{m}^3/\text{s}$ and $10\text{m}^3/\text{s}$, differ from each other and from the overall mean observation? In this case, large differences are preferred to small differences, since the latter indicates that, on average, different forecasts are followed by different observations.
- Sharpness: depicts the variability of the forecasts, as described by their marginal distribution $f(\mathbf{y})$. For example, in Fig. 1.2 we have illustrated the PDFs of two different ensembles, of which ensemble "A" is sharper (less dispersed) than ensemble "B". Sharpness and resolution coincide when the forecasts of interest are completely reliable, that is when $\bar{o}_m = m$ for all m probability thresholds.
- Consistency: expresses the degree to which ensembles contain observations identified as equiprobable members. For example, consider at time t the observation $o^t = 3$ and the following EPS, $\mathbf{y}^t = \{2, 1, 6, 14, 0, 7, 8, 7, 0, 15\}$. The observation rank within this sorted

[†]Note that mean, median, and mode coincide with the peak of the Gaussian distribution.

ensemble is equal to 5 ($\{0, 0, 1, 2, \mathbf{3}, 6, 7, 7, 8, 14, 15\}$). We must thus evaluate the observation rank within all N observation-forecast pairs. Only if the distribution of the observation rank frequency is uniform have we a completely consistent EPS.

- Uncertainty: represents the variability of the observations, as described by their PDF. A situation for which events are approximately equally likely is indicative of a relatively high uncertainty, whereas a situation for which one or two events predominate is indicative of a relatively low uncertainty.

1.3 Verification statistics for ensemble forecasts

Aiming to establish a probabilistic distribution at each time-step, the modeller must choose between parametric and nonparametric distributions.

In the meteorological community, it is commonly accepted that data is normally distributed. However, for hydrological applications, the Gamma distribution makes more sense given the asymmetry in the distribution of precipitation and discharge data [152]. But, the gamma function involves more complex evaluations than the normal distribution, which has explicit mathematical expressions. In such case, Székely [141] proposes Monte Carlo techniques for the adjustment of a non-normal distribution to the ensembles. Nonparametric distributions, such as an empirical step distribution or a kernel-based method of estimation [89, 161], offer an alternative to non-normal distributions.

It is also important to note that the robustness of the analysis of any property of an EPS is dependent on the number of cases evaluated. Ideally, we should have a large number of forecast-observation pairs. Finally, each property is always represented by a central tendency measure, such as the mean or the median.

1.3.1 Continuous ranked probability score

Continuous Ranked Probability Score (CRPS) consists of the integral of the Brier score[‡] in the continuous variable domain [79]. This score is defined as the squared error between the ensemble Cumulative Distribution Function (CDF), $F(\mathbf{y}^t)$, and the "fictitious" CDF of the observation at each time-step t. The latter is described by the Heaviside step function, which is equal to zero when forecasts are less than the observation, $H(y_i < o^t) = 0$, and equal to one otherwise, $H(y_i \geq o^t) = 1$ (Fig. 1.3):

$$\mathrm{CRPS}\left(\mathbf{y}^t, o^t\right) = \int_{-\infty}^{+\infty} \left(F(\mathbf{y}^t) - H(\mathbf{y}^t, o^t)\right)^2 dy. \tag{1.6}$$

Smaller CRPS values indicate better performance. The mean CRPS is equivalent to the Mean Absolute Error (MAE) for a deterministic forecast [79], i.e. when the step function is applied to

[‡]Brier Score measures the mean square error of the forecast probabilities where the observations are either 0 (no occurrence) or 1 (occurrence).

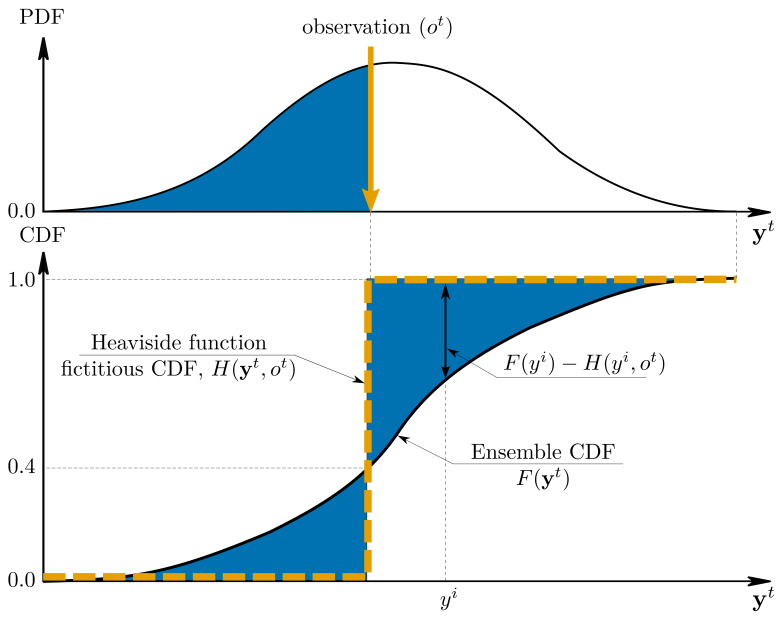


Figure 1.3: Continuous ranked probability score evaluation.

both the single forecast and the observation. The CRPS simultaneously evaluates reliability, resolution, and uncertainty [66, 79].

Assuming that the forecast ensembles $\mathbf{y}^t$ are normally distributed, the CRPS at time t is defined by [66]:

$$\begin{aligned} \mathrm{CRPS}\left(F\left(\mathbf{y}^t\right), o^t\right) \quad = \quad & \sigma^t \left[\frac{1}{\sqrt{\pi}} - 2\phi\left(\frac{o^t - \mu^t}{\sigma^t}\right)\right. \\ & \left. - \left(\frac{o^t - \mu^t}{\sigma^t}\right)\left(2\Phi\left(\frac{o^t - \mu^t}{\sigma^t}\right) - 1\right)\right], \end{aligned} \tag{1.7}$$

where ϕ and Φ denote the normalized variables for the PDF and CDF of the standard Gaussian distribution, o^t is the observation, μ^t the mean forecast, and σ^t the standard deviation.

1.3.2 Ignorance score

Proposed by Good [67] as the logarithmic score, the IGNorance Score (IGNS) is defined as the logarithm of the ensemble probability density function $f(\mathbf{y}^t)$ at the point corresponding to the observation o^t:

$$\mathrm{IGNS}\left(\mathbf{y}^t,\ o^t\right) \ = \ -\log_2\ \left[f\left(\mathbf{y}^t\right)_{o^t}\right]. \tag{1.8}$$

Note that this score can take negative values because the PDF may be larger than one[§].Smaller values indicate better performance. The IGNS is a local measure that severely penalizes the

[§]Any function $f(x)$ is potentially a PDF if its satisfies two conditions: $f(x)$ is non-negative and its integral is equal to one. Satisfying these conditions, the PDF can be greater than 1.

bias, because positioning the observation in forecast regions of low probability leads to values that tend to infinity.

Roulston and Smith [128] call attention to such situations because reporting zero forecast probabilities is difficult to justify, especially if the forecast probabilities are obtained from finite ensembles and imperfect models. Forecasters should replace zero forecast probabilities with small probabilities based on the uncertainties in the forecast PDF. In another way, following Boucher et al. [22], infinite values can be replaced by the next worst non-infinite value. This score is highly sensitive to extreme cases [66]. To rule out the possibility that the results solely reflect the effect of a few outliers, Weigend and Shi [157] proposed the trimmed mean as a measure of central tendency, excluding the highest and lowest 2% of the IGNS values.

The binary logarithm version of the IGNS measures the information deficit in bits (unit frequently used in information theory). To calculate this score in other bases, e.g. a decimal one, reduces the "sharpness" of the penalty function (Fig. 1.4). In information theory, when logarithms are calculated in base e, the unit of information by convention is a natural digit or nat, nit, or natural bel; and in base 10, a hartley, bel, digit, decimal digit, or decit [162].

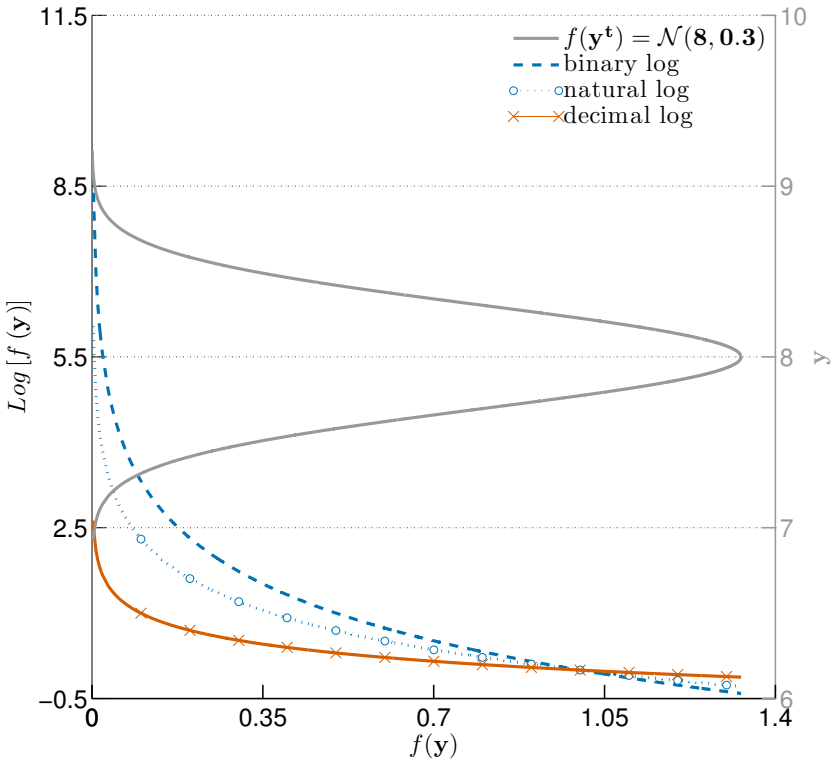


Figure 1.4: Ignorance score evaluation.

1.3.3 Reliability diagram

The Reliability Diagram (RD) or attributes diagram is a graphical representation of the joint distribution of the forecasts and observations, for probability forecasts of a binary predictand [161] (See in Section 1.2 the reliability evaluation, Fig. 1.2). For its construction, one defines M probability thresholds, often deciles, then one computes the conditional observed frequency for each of these M thresholds. Finally, one illustrates the relationship between forecast probabilities and conditional observed frequency.

In a perfectly reliable system, $\bar{o}_m$ will be equal to m, i.e. the distance or area between the 1:1 line and computed pairs $(m, \bar{o}_m)$, will be very small (left panel of Fig. 1.5). Consequently we can evaluate the reliability of the system from MSE assessed from the differences between conditional observed frequencies and evaluated probability thresholds in the RD ($\mathrm{RD_{MSE}}$):

$$\mathrm{RD_{MSE}}(\mathbf{Y},\ \mathbf{o}) \ = \ \frac{1}{M} \sum_{i=1}^{M} (\bar{o}_{m_i} \ - \ m_i)^2 \,, \text{ where: } m_i \in [0,1] \,. \tag{1.9}$$

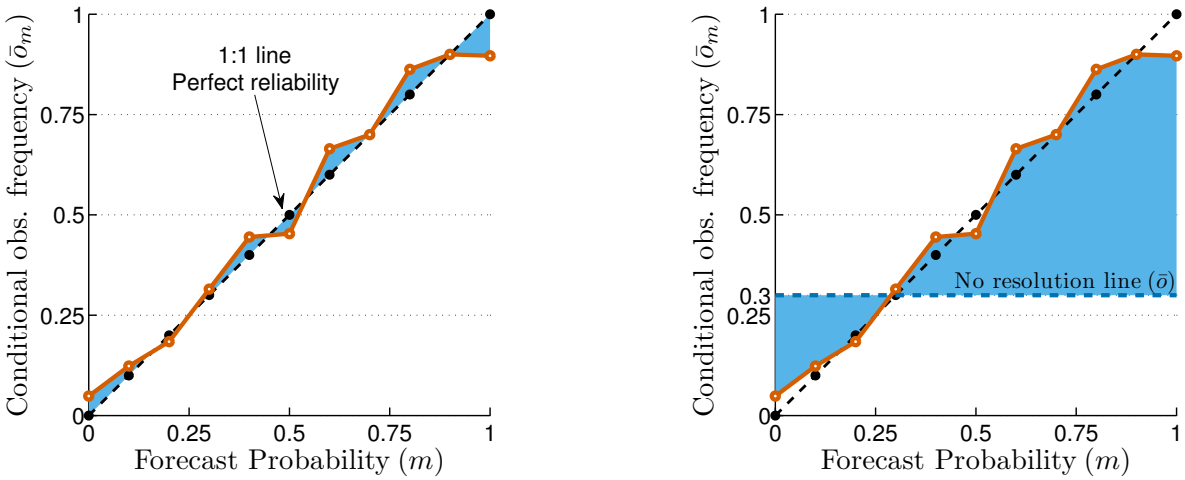


Figure 1.5: Reliability diagram.

Note that Eq. 1.9 corresponds to the reliability as defined in the Brier score decomposition [161]. It is clear that to maximize the reliability one seeks to minimize Eq. 1.9.

The reliability diagram proposes a direct assessment of reliability and resolution of a probability forecast. Regarding the resolution (ability of the forecast to distinguish situations with distinctly different frequencies of occurrence), its measure is given by the difference between each of conditional observed probabilities $\bar{o}_m$ and the overall unconditional mean observation $\bar{o}$ (see No-resolution line in right panel of Fig. 1.5).

Finally, a reliability diagram diagnosis leads to determine overforecasting or underforecasting. For example, if the curve is below the 1:1 line, that indicates that the average forecast is

larger than the average observation (overforecasting). But, if the curve is above the 1:1 line (underforecasting), the average forecast is smaller than the average observation.

1.3.4 Rank histogram

The Rank Histogram (RH) is a graphical tool that was devised independently by Anderson [15], Hamill and Colucci [76], and Talagrand et al. [142]. For its elaboration, the rank of the observation within each ensemble is first evaluated at each time-step (as the example presented in Section 1.2), then a histogram of the observation ranks is plotted (Fig. 1.6). In the case of equality of observation with one or more of the ensemble members, the observation rank is chosen randomly among them.

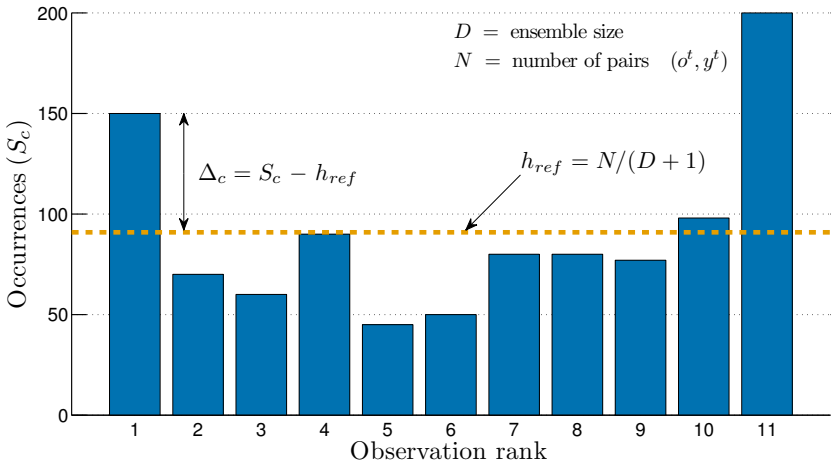


Figure 1.6: Rank histogram.

A perfectly consistent system will produce a flat histogram. For a reliable system of D forecasts, over all $D+1$ histogram bins (because the observation is added to the ensembles at each time-step), the number of elements in each bin, S_c, has an expected value $N/(D+1)$, while the deviation of the histogram from flatness, Δ, is measured by [40, 142]:

$$\Delta = \sum_{c=1}^{D+1} (S_c - h_{\text{ref}})^2 \text{, where: } h_{\text{ref}} = \frac{N}{D+1}. \tag{1.10}$$

A reliable system has an expectation of $\Delta_0 = \frac{DN}{D+1}$. The δ ratio ($\delta = \Delta/\Delta_0$), proposed by Talagrand et al. [142], is used as a scalar measure of the reliability of an ensemble prediction system. A value of δ that is considerably larger than 1 is a measure of unreliability.

Reliability, consistency and bias of the ensemble are evaluated from this score. If observations frequently fall in extreme intervals rather than in the middle ones (Fig. 1.6), it is an indication that the ensemble spread is too small (U-shaped histograms). But if the histogram is dome-shaped, it is an indication of an ensemble spread too large. Finally, an asymmetric histogram depicts a biased ensemble.

It should be stressed that consistency is directly related to reliability, although ensemble consistency does not necessarily imply that probability forecasts constructed from the ensemble are reliable, unless either the ensemble size is relatively large or the forecasts are reasonably skillful, or both [160].

1.4 Conclusion

In this chapter we presented tools to evaluate the quality of EPSs highlighting similarities between the concepts developed in the machine learning and hydrometeorological communities. Initially, we introduced some notions such as bias, variance and covariance of predictors in terms of MSE giving origin to the mathematical definition of diversity. Subsequently, we displayed the properties commonly evaluated in HEPS with a summary of the tools popularly used for such purposes.

In the first section, we presented developments in the machine learning community around the notions behind the evaluation of the MSE for different deterministic scenarios: single-predictor model, expected error for multiple experiments with the same model (a.k.a. bias-variance dilemma), multi-predictive model, and the expected error of a multi-predictor model (a.k.a. bias-variance-covariance dilemma). In the latter, we illustrated how the covariance between predictors can be theoretically manipulated to reduce the ensemble expected error. However, this deterministic view based on the error reduction of an "optimal" single forecast derived from an ensemble does not take into account other properties of ensembles such as reliability, consistency, and sharpness, which are presented in the second section.

Finally, in the third section, we exposed briefly probabilistic tools, called scores, which will be used throughout this investigation. Given the complex relationship between the various properties of these scores, the two topics discussed in this thesis will be addressed from a multi-score framework that will allow the reader to understand a little more the purpose of these tools.

Chapter 2

Machine Learning: Some Concepts and Tools

This chapter briefly describes some key concepts such as generalization, clustering, ANNs and features selection, applied in the simplification of a conventional HEPS and the HEPS construction based on ANN ensembles.

The concept of **generalization**, which refers to the capacity to predict a correct output from examples that differ from those used in the model training, is explained first. This concept is exploited under HEPS simplification in two ways: promoting data selection to apply different techniques for predictors selection and evaluating the ability of extrapolating the simplified HEPS. Referring to the construction of HEPS with ANN, it is well-known that the conservation of the generalization property is at the core of ANN training.

A classical algorithm to group information that is somehow similar (**clustering**) is presented next. In the HEPS simplification case, this technique is used in two ways: grouping basins to evaluate a regional simplification HEPS scheme and filtering precipitation forecasts from MEPS to generate representative meteorological members. In the case of ANN, clustering takes a central place in the the application of the MLD model regarding datasets selection for training each ANN forming the prediction ensemble and model inputs spaces definition to impose more variability in the HEPS.

An overview of the most popular ANN architecture in hydrology is proposed last, namely FFNN. Different techniques for **feature selection**, a typical machine learning topic which we seek to select the most relevant features for use in model construction* are presented. Again, we emphasize the adaptation of this approach throughout this thesis. The similarity with HEPS simplification, based on predictors selection, because in this case instead of selecting features we select a predictors subset for reducing HEPS complexity and even increasing the

*In machine learning and statistics, feature selection, is also known as variable selection, attribute selection, variable subset selection or input variable selection.

quality of HEPS in some cases. Finally, in the case of HEPS based on ANN ensembles, we propose an IVS based on recursive application of clustering in a classical stepwise technique called Forward Greedy Selection (FGS) to forcing ensembles diversity. More details on those concepts and tools are provided in Alpaydin [7], Bishop [20], and Duda et al. [56].

2.1 Generalization ability

The model quality or its generalization ability is defined as its capacity to predict the right output from examples that differ from those used in training. Generalization depends not only on the quantity and quality of the information but also on factors related to bias and model variance; these concepts were discussed in Sect. 1.1.2.

In hydrology, both the generalization and complexity of the models have also been debated. In this regard, Sivapalan et al. [138] show that, whereas in the 1970s and 1980s it was sometimes argued that more complex models may be appropriate, the opposite is a mathematical consequence i.e. the more complex a model the more data are needed to train and test it. Due to non-existent or inadequate data, many models of this type tend to be over-parametrized with arbitrary and overly complex model structures [19]. For its part, Kirchner [86] considers that scientific progress will mostly be achieved through the collision of theory and data, rather than through increasingly elaborate and parameter-rich models that may succeed as mathematical marionettes, dancing to match the calibration data, even if their underlying premises are unrealistic. Advancing the science of hydrology will require not only developing theories leading to the right answers but also testing whether they produce the right answers for the right reasons.

To clarify these concepts, in Fig. 2.1 we present a typical regression problem. Here, calibration[†] and validation data are taken from a parabolic model with added noise for simulating a real problem. In this example, we analyze the generalization ability of three regression models: a constant (0^{th} degree polynomial), a 2^{nd} degree polynomial (parabolic), and a 10^{th} degree polynomial. We show two trials per model in which calibration data slightly differ.

The results show that the more complex model (10^{th} order) allows a perfect fit to the calibration data ('apparent' low bias) but small changes in the dataset cause a larger change in the fitted polynomials; thus variance increases. In contrast, the 0^{th} and 2^{nd} degree models are by no means a perfect fit, but are clearly more stable (less variance). To reduce bias, the model should be flexible, at the risk of having high variance. If the variance is kept low, we may not be able to make a good fit and may end up with a large bias. The optimal model is the one that has the best trade-off between bias and variance [7]. This bias/variance dilemma is true for any data-driven system and not only for polynomial regression [65].

[†]In our context calibration and training mean the same thing.

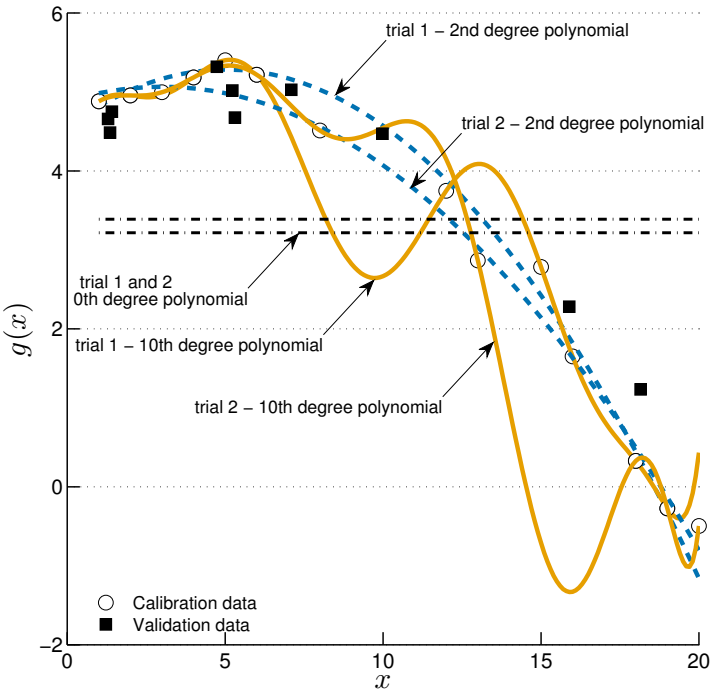


Figure 2.1: Generalization problem.

The above example illustrates how random noise in the training phase can lead to overfitting (the model learns the noise). In fact, overfitting is possible even when the calibration data are noise-free, especially when a relatively small number of examples are used for calibrating a model with a relatively large number of parameters, which is typical of ill-defined problems [106]. Fig. 2.1 illustrates such a situation for which a 11-parameter model (10^{th} degree fitting) is calibrated with only 10 examples.

A markedly accentuated bias indicates that our model does not contain the solution, which is defined as underfitting (e.g. the constant model in Fig. 2.1). In general, given models of comparable errors, a simple model would generalize better than a complex model. This principle is known as Occam's razor that states that simpler explanations are more plausible and any unnecessary complexity should be shaved off [7]. Despite this, it is important to note that there is not necessarily a relationship between the number of parameters and the tendency to over-train, for example Domingos [52] shows that model ensembles are filters with a smoothness assumption on the true function or that support vector machines can effectively have an infinite number of parameters without overfitting. Consequently, Domingos [52] proposes a modification to the parsimony principle:

> "Contrary to intuition, there is no necessary connection between the number of parameters of a model and its tendency to overfit. A more sophisticated view instead equates complexity with the size of the hypothesis space, on the basis that smaller spaces allow hypotheses to be represented by shorter codes. But viewing

> this as 'proof' of a trade-off between accuracy and simplicity is circular reasoning: we made the hypotheses we prefer simpler by design, and if they are accurate it is because our preferences are accurate, not because the hypotheses are 'simple' in the representation we chose."

To combat overfitting, the machine learning community has developed certain techniques such as regularization and early stopping [7, 74]. The latter is one of the most successful methods because of its simplicity [106]. In this technique, the training dataset is divided into two subsets: estimation and validation. The first one is exposed to some iterative search of model parameters, while the second one serves to simulate the generalization ability of the model.

The validation error normally decreases during the initial phase of parameters estimation, as does the estimation set error. However, when the model begins overfitting, the error on the validation set typically begins to rise. When the validation error increases, the optimization within the estimation dataset is stopped. The reader will note that throughout this thesis we implement actively the early stopping technique to confront overfitting in predictors selection (Part II) or in ANN ensembles construction (Part III).

This subdivision of data can also be used to assess the overall relevance of a model by a procedure known as the cross-validation process. For example, in order to find the correct order of a polynomial regression (Fig. 2.1) given a number of candidate polynomials of different orders, for each order we find the coefficients on the estimation subset, calculate their errors on the validation subset, and select the one that has the least validation error as the best model [7].

At this point, the influential role of the different datasets in the experimental design is evident. Regarding the datasets nomenclature, both the hydrological and machine learning communities have not established a general consensus. Table 2.1 provides a general overview of this terminology conundrum. We adopted here the nomenclature proposed by Haykin [77].

It is important to stress that the main source of confusion focuses on the dataset called validation. While in the hydrological community this term is used for the dataset used to evaluate the model generalization, in the machine learning community it is frequently adopted to designate the dataset employed to control the parameter optimization or to apply cross-validation. Finally, it is worth noting that a rigorous evaluation of the model's performance should be based on data that were not used in the training phase, otherwise the apparent accuracy reflects an "optimistic" model. In contrast, when a relatively large proportion of the information is used for validation, the resulting performance can be highly "pessimistic" [51]. In practice, accuracy estimates are made by constructing disjoint calibration and validation sets using sampling methods.

Table 2.1: Differences of datasets nomenclature between hydrological and machine learning communities.

Adopted here	Community		Objective
	Hydrological	Machine Learning	
Training	Calibration Optimization	Training	To establish the model. It is subdivided into estimation and validation datasets if early stopping or cross-validation is required
Estimation	Calibration Optimization	Training Optimization Estimation	To execute an iterative procedure of model parameter evaluation based on measures such as MSE
Validation	Verification Validation	Validation	To monitor error generalization in the training phase
Test	Verification Validation Test	Test Validation	To calculate the "real" model performance using information that was not used in any training stage

2.2 Clustering

Clustering involves dividing a set of data points into groups or clusters. Each group consists of objects that are similar between themselves and dissimilar to objects of other groups. Clustering techniques have traditionally been applied to unsupervised classification problems. In fact, in such unsupervised learning problems, the methods are applied to perform a partition of the input space, to discover groups of similar examples, and to develop classification labels ignoring the information about the output variable [20]. In other situations, the objective is to construct relationships between the input-output examples.

In our context, we use clustering for the HEPS simplification problem in two ways:

1. Finding regions of several basins with similar geographic location or hydrometeorological properties to integrate the simplification schemes evaluated in each of them.
2. Selecting representative meteorological members from MEPS to simplify HEPS evaluation, i.e. instead of evaluating dozens of precipitation forecast scenarios, we seek the most representatives to simplify its propagation into the hydrological models.

On the other hand, for building a HEPS with ANN ensembles, we use clustering to obtain sample datasets that are representative and diverse, which is at the core of the diversity model that we evaluate in the dataset and input system spaces showed in the MLD model (Fig. I.3). Due to its simplicity, we employ the k-means clustering method but there is a wide

range of models to perform such work such as hierarchical clustering, multivariate normal distributions used by the Expectation-Maximization algorithm, or Kohonen maps (also called self-organizing map).

2.2.1 *k*-means algorithm

Given a set of N observations $(\mathbf{x}^1, \ldots, \mathbf{x}^N)$, where each observation is a D-dimensional vector, k-means algorithm aims to classify the data set into k subsets (clusters), so that the data in each subset (ideally) share some common traits – often proximity according to some defined distance measure to the cluster centroid (m_k). The number of clusters K is specified beforehand.

For each data point $\mathbf{x}^t$, there is a corresponding set of binary indicator variables $b_k^t \in \{0,\ 1\}$, so that if data point $\mathbf{x}^n$ is assigned to cluster k then $b_k^n = 1$ and $b_j^n = 0$ for $j \neq k$.

The goal is to find values for b_k^t and the centroid of the cluster ($\mathbf{m}_k$) that minimize an objective function that represents the sum of the squares of the distances of each observation to its assigned vector ($\mathbf{m}_k$). The objective function is given by:

$$J \ = \ \sum_{t=1}^{N} \sum_{k=1}^{K} \ b_k^t \|\mathbf{x}^t \ - \ \mathbf{m}_k\|^2. \tag{2.1}$$

Adjustments are made by an iterative procedure [20]. The first step consists in initializing m_k, for example randomly. Then, the evaluation of Eq. 2.1 at each iteration is required for finding the $\mathbf{m}_k$ centres and using those centroids as references for the next partitioning of all the data points. This optimization procedure is then repeated until convergence. Algorithm 1 shows a pseudo-code for this procedure.

Algorithm 1 k-means pseudo-code

1. Define the number of clusters (K)
2. Initialize randomly centres $\mathbf{m}_k$ for $k = 1, \ldots, K$

repeat

  for all $\mathbf{x}^t$ where $t = 1, \ldots, N$ **do**

$$b_k^t \ = \ \begin{cases} 1 & \text{if } t \ = \ \text{argmin}_k \|\mathbf{x}^t \ - \ \mathbf{m}_k\| \\ 0 & \text{otherwise} \end{cases}$$

  end for

  for all $\mathbf{m}_k$ **do**

$$\mathbf{m}_k \ = \ \frac{\sum_{t=1}^{N} \ b_k^t \ \mathbf{x}^t}{b_k^t}$$

  end for

until $\mathbf{m}_k$ converges

2.3 Artificial neural networks

An Artificial Neural Network (ANN) is composed of computing units or neurons and their interconnections. Each computing unit collects the information from n inputs and integrates them resorting to a function that normally is the sum of the inputs [126]. An activation function is used next for comparing this sum with a threshold. In general, the units of an ANN are often numbered by layer, instead of following a global numbering. Each layer is composed by a set of parallel computing units. Usually all units from one layer are connected to all other units in the following layer.

The most common network topology in water resources is the three-layer FFNN, formed of an input layer, a hidden layer with a sigmoid transfer function, and a linear output layer [97, 98]. Theoretically, these ANN structures have the ability to approximate any non-linear function [81]. The adjustment of the parameters is performed through a learning algorithm such as the popular Back-propagation (BP) that allows supervised mapping between input vectors and corresponding target vectors using the method of gradient descent. Many variants of the BP algorithm have been developed for reducing the computation time and avoiding problems associated with the convergence to a non-optimum local minimum. For example, the Levenberg–Marquardt BP algorithm is known to provide a fast stable convergence, more reliable than any other BP variants [74].

In ANN, the lack of generalization is due to overfitting while the inefficacy to map the input vectors and corresponding target vectors is known as underfitting. Many approaches have been used to avoid these drawbacks, which are mainly associated with non-linear components of ANN and with the number of parameters. It is often proposed to keep the number of nodes in the hidden sigmoid layer to a strict minimum, but to avoid losses in the generalization, it is often proposed to provide the ANN with plenty of neurons in the hidden sigmoid layer and to use an early stopping criterion to halt training before complete convergence [11]. Three approaches have been commonly used for early stopping [135]: stop training when a predefined number of training iterations is reached, when a predefined error rate for the training set is reached, or when a minimum error rate is reached for a validation set. For this, and to evaluate generalization, the available data must be split into three parts: the estimation, validation, and test datasets. The estimation set is used to compute the model parameters, the validation set is used to compute the stopping criterion, and the test set is used to assess the predictive ability of the trained model (generalization).

2.4 Feature selection

In feature selection or IVS, we are interested in finding the best input subset (or features) that provide us with the most information about the whole reference dataset. The best subset contains the least number of inputs that most contribute to the accuracy of the model.

Features selection algorithms, as well as predictors selection, can roughly be grouped into two categories depending on their application model, namely *filter* and *wrapper* methods. Filter methods allow the selection to be made without involving the chosen learning/combining system, using instead some other measures, generally statistical ones such as linear correlation or mutual information.

In contrast, wrapper methods use the performance of the chosen learning/combining system to guide the selection [7]. It is generally accepted that wrapper methods lead to higher performance [88] at the expense of high computational cost compared to filter approaches. For example, popular stepwise methods such as Forward or Backward Greedy Selection techniques add or remove or add variables one by one, so they are frequently categorized as a "brute force" method, although it is not necessarily so [73]. Meta-heuristic procedures such as EA have been proposed [58, 147, 166] as a wrapper method for features selection, with the advantage of reducing the computational cost, but also showing a capacity to find better solutions given its global search capabilities. Techniques used in this study follow.

2.4.1 Negative correlation maximization

As discussed in Sect. 1.1.3, the manipulation of the negative correlation between predictors is related to the MSE reduction, therefore this property is the basis of some training methods of prediction ensembles such as ANN ensembles trained on the negative correlation [37]. Consequently, a filter method, usually used as a pre-processing step because they are simple and fast, involves checking the pairwise correlation between the variables in order to attenuate it, i.e. their high redundancy.

2.4.2 Mutual information maximization

Mutual information maximization, a filter technique in feature selection, considers key aspects such as the non-linear relationship between features (mutual information), the degree of correlation between the features themselves (redundancy), and the consideration of the dependent variable (targets in regression problems) as an indicator of the relevance of each feature. It maximizes the first parametrized order of the utility criterion given by Eq. 2.2 – assuming that only conditional and unconditional pairwise relations exist and no higher order relations, i.e. the evaluation of mutual information (I) is truncated to the second order. This measure represents a trade-off between the individual predictive power of the feature (relevance), the unconditional correlations (redundancy), and the class-conditional correlations (conditional redundancy):

$$J_k = \underbrace{I(\mathbf{y}_k;\mathbf{o})}_{\text{relevance}} - \beta \underbrace{\sum_{m=1}^{nv} I(\mathbf{y}_i;\mathbf{y}_m)}_{\text{redundancy}} + \gamma \underbrace{\sum_{m=1}^{nv} I(\mathbf{y}_k;\mathbf{y}_m|\mathbf{o})}_{\text{conditional redundancy}}, \tag{2.2}$$

where $\mathbf{y}_k$ represents the evaluated feature, $\mathbf{o}$ the dependent variable, nv the number of previously selected variables, and β and γ the configurable parameters that must be set experimentally. Several authors have proposed different criteria with various penalties to manage the redundancy. Brown [37] presents an overview of different strategies proposed in the literature.

Evaluation of the terms of Eq. 2.2 requires also the discretization of the information. For this reason, in HEPS simplification, we use this selection technique in a hybrid fashion, optimizing both the number of classes of the discretization task and the mutual information conceptualization, based on a linear search to minimize the combination of different scores that evaluates different HEPS properties.

2.4.3 Stepwise selection

This wrapper technique adds (FGS) or removes (BGS) variables sequentially, one at a time. FGS for example involves starting with no variable and adding variables (one at each step) that improve the performance most, until any further addition does not decrease the error (a threshold gain may be put in place). In contrast, BGS starts with all d candidate variables and remove them one by one, at each step removing the one that decreases the error the most (or increases it only slightly), until any further removal increases the error substantially [7].

The BGS proceeds as follows:

1. It begins with a subdivision of the whole or reference dataset ($\mathbf{G}^d = \{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, ..., \mathbf{y}_d\}$), containing all of the original "d" candidate variables into estimation (χ_t), validation (χ_v), and test (χ_p) data subsets. In HEPS simplification, the candidate variables represent all the evaluated predictors ensemble. By contrast, in HEPS based on ANN ensembles, the candidate variables correspond to the variables considered potentially important for the construction of the prediction model.
2. In the estimation set ($\mathbf{G}^d|\chi_t$), we begin the iterative process. So, in each iteration (iter $= d-1, d-2, \ldots,$ nmin) the feature or variable "$\mathbf{y}_j$" corresponds to the one that, when it is removed, has the greatest impact on the estimation set error $\mathcal{E}$ (i.e. minimizes the estimation error the most):

$$\mathbf{y}_j = \underset{\mathbf{y}_i \in \mathbf{G}^{\mathrm{iter}+1}}{\operatorname{argmin}} \; \mathcal{E}\left(\mathbf{G}^{\mathrm{iter}+1} \backslash \{\mathbf{y}_i\} \,|\chi_t\right).$$

   It is important to note that $\mathcal{E}$ must be a scalar or single value. The reference set is then updated by removing the $\mathbf{y}_j$ variable in $\mathbf{G}$.

$$\mathbf{G}^{\mathrm{iter}} = \mathbf{G}^{\mathrm{iter}+1} \backslash \mathbf{y}_j.$$

3. At this point, the error $\mathcal{E}$ in the validation set χ_v, excluding the $\mathbf{y}_j$ variable, is evaluated.

$$\mathcal{E}_v^t = \mathcal{E}\left(\mathbf{G}^t \,|\chi_\mathrm{v}\right).$$

4. The subset $\mathbf{G}^{\text{nmin}}$ of the selected variables is achieved, then the whole selection process is analyzed on the estimation and validation results.

BGS is a local search procedure that does not guarantee finding the optimal subset. For example, $\mathbf{y}_x$ and $\mathbf{y}_p$ by themselves may not be pertinent but together they may decrease the error substantially. But, because the algorithm is greedy and removes variables one by one, it may not be able to detect this.

A BGS will be evaluated in the HEPS simplification problem (Part II), while a FGS will be employed in HEPS based on ANN ensembles (Part. III).

2.4.4 Nondominated Sorting Genetic Algorithm II

Genetic Algorithms (GAs) or more generally EA are inspired from genetic coding in biology, where solutions to a system or a problem are represented into coded strings (e.g. binary, integer, real or gray coded string). The search for a global solution is regulated by rules based on Darwin's theory of the survival of the fittest, by which strings are allowed to survive from one generation (i.e. iteration) to another and to trade part of their genetic material with other strings depending on their robustness as defined by one or several objective functions. Genetic algorithms have been regularly employed for applications in water resources and hydrology [12]. More details about EA such as: individual representations, mutation and recombination operators, population models, parent selection, and survival selection can be found in Eiben and Smith [58].

In a feature selection context, the binary or integer string codification is useful in GA representation. For example, Anctil et al. [12] proposed a binary coding to orient the rain gauge combinatorial problem toward improved forecasting performance. In a multi-objective framework, where the quality of a solution is defined by its performance in relation to several objectives, Deb et al. [50] proposed the NSGA-II.

The NSGA-II uses a GA for population evolution, in combination with a fast non-dominated sorting approach to classify solutions according to the level of non-domination, and a crowding distance operator to preserve the solution diversity. The basic steps of the algorithm can be summarized as follows:

- A population of parents P_t of size N and a population of offspring Q_t of size N are assembled to form a population ($R_t = P_t \cup Q_t$). This assembly ensures elitism.
- Population R_t is then sorted according to a non-dominance criterion to identify different fronts F_1, F_2, etc.
- The best individuals will form the first front. A new parent population P_{t+1} is generated by adding the entire fronts as they do not exceed N.
- If the number of individuals present in P_{t+1} is less than N, a crowding procedure is applied on the first front following F_i, not included in P_{t+1}. The purpose of this operator is to insert

the $N - P_{t+1}$ best individuals lacking in population P_{t+1}. Individuals on this front are used for computing the crowding distance between two neighbouring solutions.

- Once individuals in the population P_{t+1} are identified, a new child population Q_{t+1} is created by selection, recombination (crossover), and mutation.
- Parent selection uses a modified tournament operator that considers first dominance rank, then crowding distance.
- The process continues from one generation to the next, until a stopping criterion halts it.

In Chap. 5, we will describe a detailed application of this technique in a HEPS simplification based on a predictor selection scheme.

2.5 Conclusion

In this chapter, we presented some concepts and machine learning techniques to be applied to the two topics investigated in this thesis: HEPS simplification and HEPS construction with ANN. In this order, generalization is presented as a key element in the development of the methodologies proposed along this study.

Similarly, we introduced the idea of clustering, a concept frequently used in conjunction with the proposed approaches. Additionally, we presented briefly some tools typically used in the problem known in the machine learning community as feature selection or IVS, which will be adapted to the problem of selection of predictors as the basis of a HEPS simplification. The methodological details of such techniques will be addressed in the chapters concerned. Also, in the case of ANN, this chapter only introduces its operation since the configuration details in the construction of the HEPS will be discussed in the respective chapters.

Part II

HEPS Simplification

Chapter 3

Optimization Criteria

We know that Hydrological Ensemble Prediction System (HEPS) obtained through the forcing of several rainfall-runoff models with Meteorological Ensemble Prediction System (MEPS) may easily reach several hundreds of scenarios at each time step, which becomes an operational burden.

In this chapter, we assess the degree of simplification, i.e. the reduction of the number of hydrological scenarios or members, that can be achieved for an 800-member HEPS configured using 16 lumped hydrological models driven by the 50 weather ensemble forecasts from the European Centre for Medium-range Weather Forecasts (ECMWF) EPS. Note that the simplification approaches explored herein are applicable to any HEPS, regardless of their nature.

Here, Backward Greedy Selection (BGS) is proposed to assess a Hydrological Models Participation (HMP) scheme, i.e. to identify the number of scenarios corresponding to each hydrological model within a subset that offers similar or better performance than a reference set of 800 hydrological scenarios, which would issue real-time forecasts in a relatively short computational time.

The methodology proposed uses a variation of the k-fold cross-validation, allowing an optimal use of the information, and employs a multi-criterion framework that represents the combination of resolution, reliability, consistency, and diversity. However, in this chapter, we focus on an analysis of scores in the BGS process. Results show that the degree of reduction of members can be established in terms of maximum number of members required (complexity of the HEPS or the maximization of the relationship between the different scores (performance).

3.1 Review of HEPS simplification

The competency of probabilistic forecast to encompass the many sources of uncertainty in HEPS has already been demonstrated [127, 129, 148]. Yet the simultaneous consideration of the uncertainty associated with both the meteorological inputs and the structural and para-

metric configuration of the hydrological models can lead to systems consisting of too many members to be computationally and operationally implementable. Note that each meteorological member represents a distinct meteorological scenario in the MEPS. Analogously, each hydrological member identifies a distinct hydrological scenario resulting of propagation of a meteorological member into a hydrological model.

On the other hand, combining information derived from the many MEPSs is an avenue that has been shown to improve early flood warning systems [78] – **T**he Observing System Research and Predictability Experiment (THORPEX) **I**nteractive **G**rand **G**lobal **E**nsemble (TIGGE) [23] favours this new opportunity. Moreover, if the parametric uncertainty of hydrological models is assessed under the principle of equifinality [19] and if the structural uncertainty is tackled through a multi-model approach, the number of scenarios in the uncertainty cascade model may rapidly turn out to be quite large. Simplification of such a HEPS thus becomes a mandatory step from an operational standpoint.

In such a context, the hydrological and meteorological communities have focused their efforts on many lines of simplification. For instance, Pappenberger et al. [115] evaluated 10-day ahead rainfall forecasts, consisting of one deterministic, one control, and 50 ensemble forecasts, as input to a rainfall-runoff model (LisFlood), for which parameter uncertainty was represented by six different parameter sets identified through a Generalized Likelihood Uncertainty Estimation (GLUE) analysis and functional hydrograph classification. Raftery et al. [123] proposed the Bayesian Model Average (BMA) methodology as a means for the statistical post-processing of the forecast ensembles derived from numerical weather prediction models. The BMA predictive PDF is a weighted average of the PDFs centred on the bias-corrected forecasts from a set of different models. The weights assigned to each model reflect that model's contribution to the forecasting skill over a training period [155]. In that line, Vrugt et al. [152] proposed evaluating BMA weights with the DiffeRential Evolution Adaptive Metropolis (DREAM) Markov Chain Monte Carlo (MCMC) algorithm.

Other studies identified the meteorological forecasts as the most uncertain component of the cascade model [83, 115, 145], triggering interest in novel member selection techniques. For example, Marsigli et al. [100], Molteni et al. [109] and Jaun et al. [83] selected MEPS members based on lagging ensembles and derived representative members through hierarchical clustering over the domain of interest. Ebert et al. [57] analyzed the relationship between the atmospheric circulation patterns and extreme streamflows to select representative members of MEPS. Finally, Xuan et al. [164] established, in a deterministic way ("best match" approach), the location of the forecast that is the most similar to the rainfall pattern of the catchment.

Cloke and Pappenberger [45] have already highlighted the computational demand of using MEPS for flood forecasting as one of the main points to overcome in the future, either by new technologies (stochastic chip technology) or by efficient use of computing clusters. Thus,

the selection of hydrological members as part of a simplified model can be useful given the computational cost of running models and creating ensembles.

Another aspect of particular interest in the evaluation of probabilistic forecast, and therefore in hydrological member selection, is the identification of a pertinent criteria set. In conventional forecasting, i.e. when confronting an observation against a single prediction, it is now generally accepted that the calibration of hydrological models should be approached as a multi-objective problem [46, 69, 71, 156, 167]. Probabilistic forecasting is not different in that regard. In fact, the complexities of confronting an observation against an ensemble of predictions calls for a variety of criteria, here called scores, that specifically focus on one or more characteristics of the probabilistic sets. So, to assess these properties, several statistical measures should be considered concurrently [45, 161].

Few studies have experimented hydrological member selection from a multi-score point of view. Vrugt et al. [155] posed the BMA inverse problem in a multi-objective framework, examining the Pareto set of solutions between the CRPS, the MAE, and the IGNS with a method called A Multi-ALgorithm, Genetically Adaptive Multiobjective (AMALGAM) [150].

3.2 HEPS of reference

3.2.1 Configuration and catchment locations

The HEPS under study is formed of 16 lumped hydrological models forced by the 50 meteorological scenarios of the ECMWF EPS, leading to a grand ensemble of 800 hydrological members. The MEPS members are a priori assumed to be equally likely [68]. Another important feature of the HEPS at hand is the short duration of the series, from March 2005 to July 2006. This has been highlighted by several authors as a negative point in the system evaluation in the case of extreme events [45, 125].

However, other studies that focused on periods of analysis very similar to the one used here have also proven the usefulness of the ECMWF EPS. For example Rousset et al. [129] evaluated hundreds of French catchments from the 4$^{\text{th}}$ of September 2004 to the 31$^{\text{st}}$ of July 2005 showing that the information given by the ensemble forecast is useful for flood warning and water management agencies. Similarly, Thirel et al. [144], in a comparative analysis of short-range meteorological forecasts from the ECMWF EPS and PEARP EPS of Météo-France under the scheme of SIM coupling, analyzed the competence jurisdiction of each of the two MEPSs from March 11$^{\text{th}}$ 2005 to September 30$^{\text{th}}$ 2006, showing that the ECMWF EPS seemed best suited for low streamflows and large basins while the PEARP EPS was best suited for floods and small basins.

The sixteen hydrological models are lumped models and correspond to various conceptualizations of the rainfall-runoff transformation at the catchment scale. Some original model

structures were modified. Thus, to avoid unfair comparisons of models, they will be referred to hereafter as HM## (Table 3.1). It is beyond the scope of this chapter to present these models. References with a explanation of each model can be found in Velázquez et al. [148].

Table 3.1: Hydrological models and their number of parameters.

Hydrological models	Base model and # of parameters		Hydrological models	Base model and # of parameters	
HM01	CEQU	9	HM09	CREC	8
HM02	GR3J	3	HM10	GR4J	4
HM03	HBV0	9	HM11	SIMH	8
HM04	IHAC	6	HM12	MOHY	7
HM05	MORD	6	HM13	PDM0	8
HM06	SACR	13	HM14	HYM0	5
HM07	SMAR	9	HM15	TANK	10
HM08	TOPM	8	HM16	WAGE	8

It is important to note that this study focuses on evaluating the probabilistic hydrological forecasting from a cooperative point of view seeking diversity in the final hydrological members' selection, i.e. that each member acts as a complement to the others. This clarification is relevant in order to avoid misinterpretation of competitiveness in the different conceptualizations of the sixteen hydrological models used. It should be clear that the comparison would not be fair because some models such as the GR4J were specifically devised for the catchment scale, whereas others have suffered a series of changes bringing them to a lumped state.

Temperature, rainfall, and streamflow data are available at a daily time step over the period extending from 1970 to 2005, and were used for the calibration and validation of the hydrological models. Observed data for the period March 11$^{\text{th}}$ 2005 to July 31$^{\text{st}}$ 2006 was used only for the evaluation of the forecasts. The forecast validation period is thus independent of the calibration period. Rainfall data come from the meteorological analysis system SAFRAN of Météo-France (see [122] for details). They consist of rainfall accumulated at a daily time step and available for the entire country of France on an 8×8-km grid. Daily streamflow data come from the French database Banque Hydro*. The duration of available observed streamflow time series varies according to the catchment, with, on average, 29 years of available daily data for the catchment dataset used here.

The 50 perturbed forecasts from ECMWF EPS were provided at a $0.5° \times 0.5°$ lat/lon grid resolution. A detailed description of the ECMWF EPS model can be found in Molteni et al. [108] or Buizza [39]. Forecasts are issued at 12:00 UTC and extend over 240 h. Rainfall amounts were accumulated at 24 h time steps, starting at 0 h to match with observed daily data, which resulted in nine FTHs. No bias removal or disaggregation was performed. For

*The database can be found at `http://www.hydro.eaufrance.fr/`.

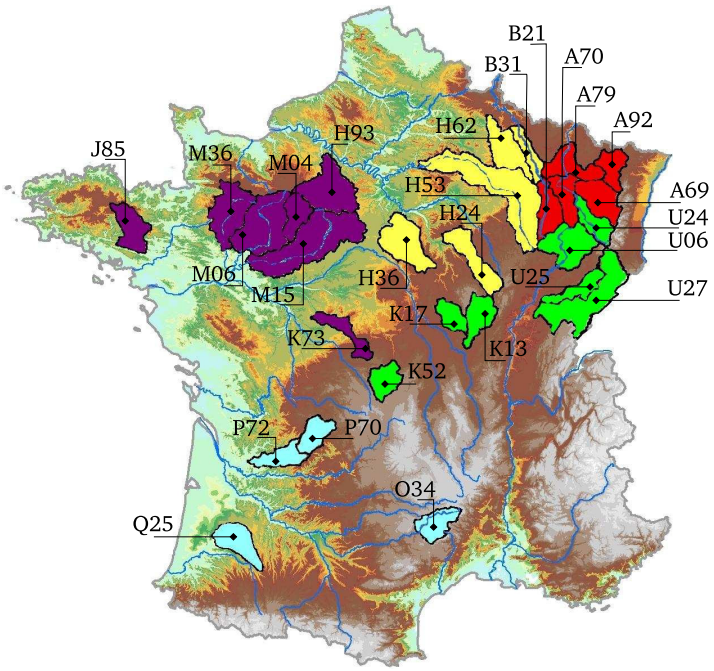


Figure 3.1: Catchment locations in the simplification schemes.

each basin, areal mean rainfall forecasts were computed by averaging the rainfall amounts of each grid above the basin, weighted by the percentage of the catchment area inside the grid.

This HEPS was implemented over 28 French catchments with an average response time of 3.2 days, representing a large range of hydro-climatic conditions (Fig. 3.1 and Table 3.2). The HEPS simplification was applied to 16 of the 28 available basins; however, for space considerations, we present only the results for 10 of them. In Chap. 4, the simplification is generalized to the other 12 neighbouring basins, highlighted in bold in Table 3.2. Hereafter, each basin is identified only with the first three characters of each code used in Table 3.2. For the distinction of the basins used in this training phase (simplified HEPS evaluation) and testing phase (generalization evaluation, Chap. 4), the latter are highlighted in bold.

3.2.2 Results of the 800-member HEPS

Figure 3.2 shows the HEPS behaviour with different set-up and different FTHs. Results focus on the reliability (RD_{MSE}) and the ensemble consistency (δ ratio) for two schemes formed from sixteen hydrological models, one led by the deterministic ECMWF EPS forecast (16-member HEPS) and the other by the 50 perturbed members from the ECMWF EPS (800-member HEPS). Results in Fig. 3.2, expressed in terms of Interquartile range (iqr) and median, are based on the grouping of the scores obtained in the 28 basins evaluated here. Note that the δ ratio and RD_{MSE} scores are directly comparable since their scale is independent of the measured variable.

Table 3.2: Main characteristics of the studied catchments (mean annual values) based on the 36-year duration of the series (1970–2006).

Catchment codes	Area (km^2)	P (mm)	ET (mm)	Q (mm)	Catchment codes	Area (km^2)	P (mm)	ET (mm)	Q (mm)
A6921010	2780	3.03	1.79	1.16	Q2593310	2500	2.52	2.25	0.73
A7930610	9837	2.77	1.80	1.20	U2542010	4970	3.63	1.76	1.88
A9221010	1760	2.47	1.84	0.87	**A7010610**	6830	2.98	1.78	1.45
B2130010	2290	2.57	1.80	0.89	**H6221010**	2940	2.49	1.84	0.91
B3150020	3904	2.57	1.80	1.08	**H9331010**	4598	1.81	1.87	0.35
H2482010	2982	2.31	1.90	0.85	**K1341810**	2277	2.65	1.89	1.04
H3621010	3900	1.97	1.96	0.45	**K5220910**	1836	2.45	1.90	0.89
H5321010	8818	2.40	1.85	0.92	**M1531610**	7920	1.85	1.95	0.35
J8502310	2465	2.35	1.90	0.81	**M3600910**	3910	2.31	1.89	0.78
K1773010	1465	2.64	1.95	1.03	**P7001510**	1863	2.87	2.09	1.18
K7312610	1712	2.13	2.01	0.67	**P7261510**	3752	2.65	2.14	0.86
M0421510	1890	2.04	1.90	0.61	**U0610010**	3740	2.93	1.85	1.34
M0680610	7380	2.04	1.94	0.56	**U2402010**	3420	3.80	1.70	2.00
O3401010	2170	3.18	1.81	1.88	**U2722010**	7290	3.63	1.79	2.06

P: precipitation, ET: potential evapotranspiration, Q: streamflow.

Figure 3.2 illustrates that the 800-member HEPS advantages become apparent and progressive after the 4th FTH. According to Velázquez et al. [148], the firsts FTHs present a low performance partly inherited from the meteorological ensembles, which are not reliable prior to about a 3rd FTH.

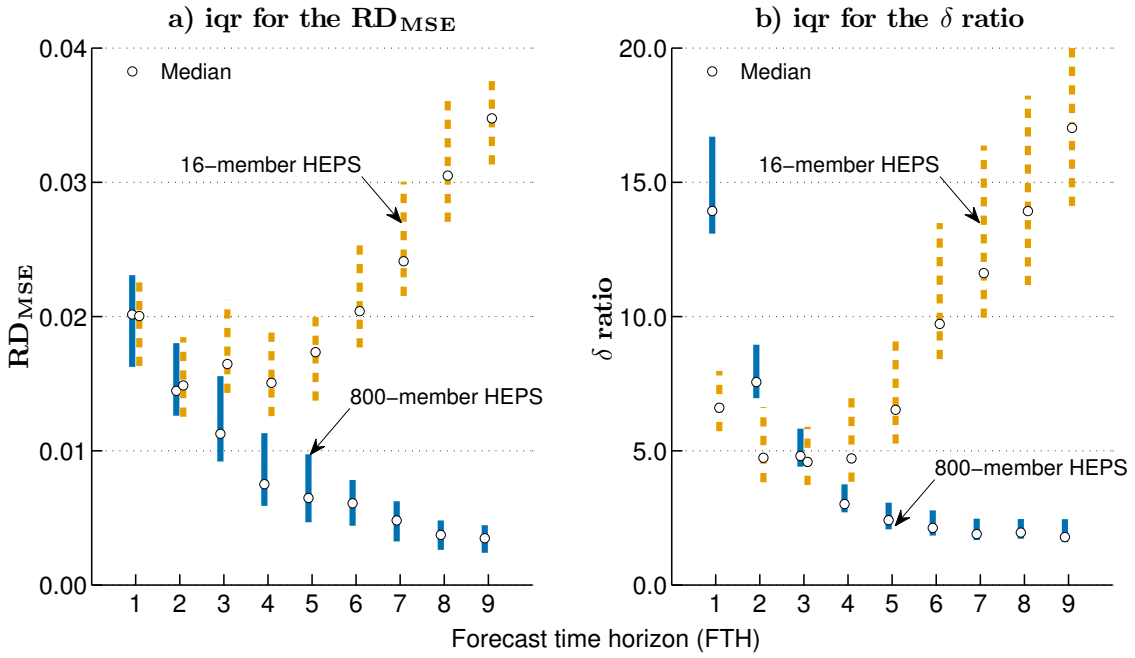


Figure 3.2: iqr of RD_{MSE} and δ ratio assessed over the 28 catchments under two HEPS schemes: the deterministic (16-member HEPS) and the probabilistic (800-member HEPS) schemes.

Taking into account the direct relationship between dispersion and reliability in this database, we consider an unitless measure: the coefficient of variation – that relates the standard deviation and the mean. First, the ensemble's coefficient of variation is calculated at each time-step, and then the MeDian of the Coefficients of Variation (MDCV) is evaluated. The first FTH presented a MDCV of 0.05 while longer FTHs reported larger values, e.g. 9$^{\text{th}}$ FTH HEPS reached a MDCV of 0.57.

Table 3.3 presents a comparison for the two HEPS schemes analyzed by Velázquez et al. [148]. It should be stressed that the 800-member HEPS serves as a reference for the selection of hydrological members since their scores confirmed their superiority over the latter FTHs.

With respect to the IGNS score, mean values are generally negative, which shows that, on average, the system has an acceptable bias. Finally, in terms of CRPS, Velázquez et al. [148] show in detail the efficiency of CRPS in this 800-member HEPS.

Table 3.3: Performance of the 16-member HEPS and the 800-member HEPS on the 9$^{\text{th}}$ FTH. Hereafter, RD_{MSE} values are expressed on a 10^{-3} basis.

Basin	HEPS (members)	CRPS	IGNS	RD_{MSE}	δ	MDCV
A79	16	0.338	4.51	93.95	42.5	0.18
	800	0.263	0.44	5.06	3.3	0.41
B31	16	0.164	0.77	39.21	21.3	0.13
	800	0.135	−0.88	4.51	2.7	0.22
J85	16	0.184	0.69	34.49	15.8	0.20
	800	0.163	−0.98	2.16	1.6	0.37
M04	16	0.177	0.49	27.24	13.7	0.19
	800	0.160	−0.99	1.74	1.5	0.37
Q25	16	0.186	0.66	32.89	14.9	0.21
	800	0.163	−0.98	2.15	1.5	0.37
B21	16	0.282	1.05	39.29	23.3	0.32
	800	0.230	−0.29	2.43	2.2	0.57
H36	16	0.181	0.84	34.89	17.4	0.19
	800	0.161	−0.99	3.50	1.5	0.37
K73	16	0.184	0.53	33.98	15.8	0.19
	800	0.165	−0.93	3.09	1.9	0.35
O34	16	0.198	0.77	36.39	16.8	0.19
	800	0.169	−0.86	3.46	1.5	0.36
U25	16	0.390	3.29	39.73	21.0	0.19
	800	0.289	−0.36	3.39	2.6	0.35

Velázquez et al. [148] have also shown the high performance of the 800-member HEPS for the 9th FTH. However, as one of the objectives is to show the transferability of the hydrological members selections to other FTHs, it is necessary to show the performance of the 800-member HEPS in such scenarios to clearly establish our point of reference concerning the quality of the hydrological members' selection.

3.3 Hydrological models participation

The simplification of the 800-member HEPS should be viewed as a direct systematic selection of certain hydrological members, which indirectly leads us to determine the HMP as the key concept of the simplification task.

To understand the HMP concept, consider the example given in Table 3.4 for a simplified HEPS of 30 members presented in the first column. Assuming that this simplified scheme provides at least a performance equal with the 800-member HEPS, Table 3.4 shows that the evaluation and posterior combination of hydrological models can be reduced substantially (7 instead of 16). Note that the last two columns show the apparent higher relevance of models 1 and 16; however, other models of less weight can be important to describe the uncertainty of the process with opposing views to those of the most relevant models.

Table 3.4: Hypothetical example to show the HMP as a key concept of the proposed simplification scheme.

30-member HEPS	800-member HEPS		Hydrological Models Participation (HMP)	Hydrological Model weight (%)
Hydrological members selected	Hydrological Model (HM#i)	Members interval $[50i-49, 50i]$		
	1	1 – 50	7	23.3
10, 12, 23, 25, 34, 42, 45, 55,	2	51 – 100	3	10.0
63, 70, 245, 247, 345, 350,	5	201 – 250	2	6.7
654, 680, 690, 700, 701, 710,	7	301 – 350	2	6.7
751, 753, 755, 757, 759, 760,	14	651 – 700	4	13.3
778, 780, 785, 800	15	701 – 750	2	6.7
	16	751 – 800	10	33.3

So, using the HMP information, one can expect that, propagating seven meteorological members through hydrological model HM#1, three for HM#2, two for HM#5, and so on to propagate ten for HM#16, we obtain a 30-member HEPS of at least the same performance as the 800-member HEPS. Now, another question concerns how do we identify the best meteorological members to propagate? Accordingly, in this and next chapter, we show the results of a random meteorological members propagation, whereas Chap. 5 shows the results of the so-called meteorological representative members.

Using the 800 scenarios of the HEPS of reference as independent variables, considering that ECMWF EPS members are equiprobable and interchangeable, raise the question about the structural coherence of such HEPS in the time domain. That is, to what extent can we consider that hydrological members represent "homogeneous" conditions of "measuring" in the time domain. Or what is the relevance of using the member's numbering in a simplification methodology? This question leads a problem that is probably more complex to solve than the object itself of this research: How does one assess the probability propagation of meteorological predictions or members under a multi-model hydrological scheme?

Note that while meteorological members are interchangeable, the occurrence of each hydrological model within HEPS stays invariable. For all time steps, the first 50 hydrological members correspond to the combination of 50 meteorological members and hydrological model #1. Similarly, the last 50 hydrological members (751-800) correspond to the same combination with hydrological model #16. It is clear that hydrological models act as non-linear filters in which one of their variables is precipitation. We assume that the hydrological models that form the HEPS of reference reflect different conceptualizations of the hydrological process. In this way, the removal of a member only represents a loss of model weight in the simplified scheme based on the HMP.

In conclusion, we hypothesize that considering each hydrological member as a variable is not in conflict with the proposed methodology, because the selection of members, for subsequent interpretation in terms of HMP, is not made on members of the ECMWF EPS but rather on the 800-member hydrological response. The empirical validity of these assumptions will be evaluated particularly in Sect. 3.5.3. On the other hand, in Sect. 3.2.2 we showed that the 800-member HEPS has a high performance on the 9^{th} FTH. Consequently, we apply the simplification process on the database corresponding to this lead time. This decision about the FTH is justified by the fact that the HMP as a method of simplifying HEPS should be unique regardless of the FTH.

3.4 Estimation of hydrological models participation

Figure 3.3 shows the proposed scheme for hydrological models participation applied to the 800-member HEPS on the 9^{th} FTH.

3.4.1 Resampling technique

In some algorithms, such as the BGS, the overfitting[†] is highlighted as a structural problem. So, we use an early stopping technique (see Sect. 2.1) for improving generalization.

[†]When the error on the training set is driven to small values, but the error of the model is large on new data.

The need to define three subsets to run the BGS and the short duration of the series impose the use of resampling techniques such as k-fold cross-validation, which maximizes the utilization of the available information.

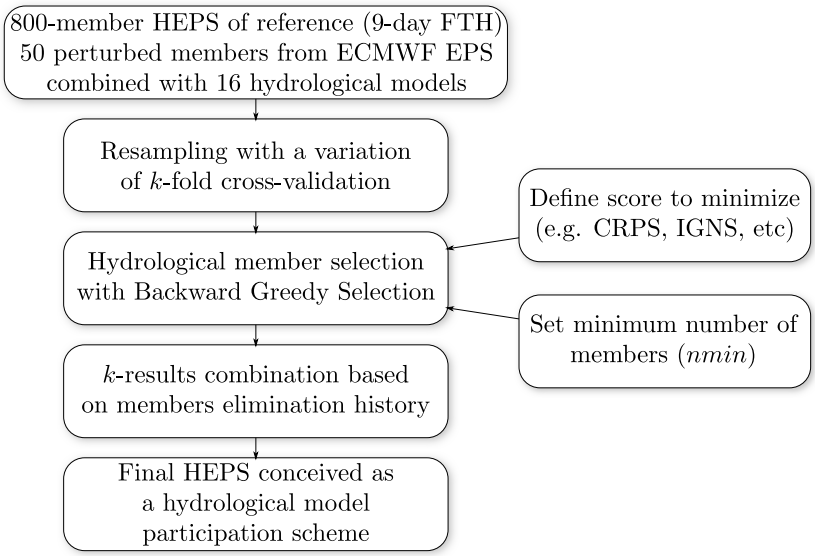


Figure 3.3: Evaluation of HMP.

Moreover, one notes the high degree of linear correlation exhibited in the first lags of the correlogram of the streamflow series at hand (e.g. in 80 % of catchments evaluated, the correlation using a lag of three days was greater than 0.82). So, the choice of the estimation and validation data should be careful. For example, suppose that the linear correlation between o^t and o^{t+1} is equal to 0.8 and that the selection of hydrological members has been trained in o^t and validated in o^{t+1}. The validation could consequently be highly contaminated by the effect of the correlation between data. Correlation contamination is avoided by forming estimation and validation subsets from groups of 10 consecutive data (blocks) rather than from individual data. It is important to note that contrarily to standard hydrology applications, the order of the events is not important in the BGS process.

Here, the dataset is divided into 5 equal-sized parts in order to create 5 experiments. In each experiment, a part is kept out for testing, while the remaining four parts, a priori divided in blocks, are randomly combined to form estimation and validation subsets. The detailed process develops in two steps:

- *Step 1: Data and test set configuration.* The test set is set-up from simple cut-offs to "guarantee" statistical independence with the estimation-validation process. To build the test set, the series is subdivided into five folds, each of which corresponds to the test set of each experiment. For example, if N denotes the length of the series, the test set of the first experiment corresponds to the first fold ($i = 1$ to $\lfloor N/5 \rfloor$), similarly the test set of the fifth experiment will be the last fold ($i = \lceil 4N/5 \rceil$ to N). Thus, strong linear correlation between

estimation-validation and the test dataset is limited only to the values situated near the cut-off line.

- *Step 2: Block selection of the estimation and validation sets.* The remaining 4 parts are grouped into k blocks of consecutive pairs of observation-ensemble forecast, then we randomly choose 75 % of the blocks for the estimation set and the remaining 25 % sets for the validation set.

3.4.2 Backward greedy selection - Setup

In Machine Learning, the evaluation of multiple models for simulation or prediction of an event, and to further select those which together enhance or simplify a condition for adjustment, is known as an "overproduce and select" procedure. In a general context of selection, numerous methods have been developed (see Sect. 2.4).

Here, BGS (see Sect. 2.4.3) and the idea of subdividing the data into three subsets to improve the generalization are applied. The mechanism of hydrological member elimination begins with all members, removing at each step the hydrological member that, when it is removed, has the greatest impact on the estimation set error (i.e. minimizes estimation error the most). For its implementation, it is necessary to define the error function (score) and the minimum number of members.

Score to minimize

Since our main interest is to evaluate the relationship between probabilistic properties, we independently use six error functions: the four scores defined in Sect. 1.3 (CRPS, IGNS, MSE on reliability diagram, and rank histogram flatness), the MDCV, and a function that combines all previous ones, called the Combined Criterion (CC):

$$\mathrm{CC} \;=\; w_1 \frac{\overline{\mathrm{CRPS}}_{\mathrm{se}}}{\overline{\mathrm{CRPS}}_{\mathrm{ie}}} + w_2 \frac{z_1 - \overline{\mathrm{IGNS}}_{\mathrm{se}}}{z_1 - \overline{\mathrm{IGNS}}_{\mathrm{ie}}} + w_3 \frac{\mathrm{RD}_{\mathrm{MSE}_{\mathrm{se}}}}{\mathrm{RD}_{\mathrm{MSE}_{\mathrm{ie}}}} + w_4 \frac{\delta_{\mathrm{se}}}{\delta_{\mathrm{ie}}} + w_5 \frac{z_2 - \mathrm{MDCV}_{\mathrm{se}}}{z_2 - \mathrm{MDCV}_{\mathrm{ie}}}, \qquad (3.1)$$

where se and ie subscripts represent the selected ensemble of hydrological members and the initial 800-member ensemble, respectively. The weights (w_i) offer the possibility of constructing a trade-off among different objectives. The threshold z_1 manipulates the duality of having a positive (or negative) IGNS in the selection ensemble as in the 800-member set. The threshold z_2 is used to change the MDCV orientation since the objective is to maximize dispersion, given the low variability of the 800-member HEPS under study in many cases.

The latter is proposed because selecting only one criterion may give a partial view of the forecast performance thus be misleading. The combination of several metrics into one diagram has already been evaluated [143]. But it is inappropriate for this study because a scalar objective value is required for the selection procedure. So, we propose the following guidelines to define the CC:

- The combination should assign weights to each of the scores as a direct measure prioritizing some of the characteristics of the HEPS in evaluation. In our case, weights were used only to give priority to the reliability in the selection, because Velázquez et al. [148] showed that this was the most influential aspect in the evaluation of the HEPS studied here. For this reason, the weight assigned to the reliability (RD_{MSE}) corresponds to twice that of the other factors, which have a unit weight.
- Each score in the selected ensemble of hydrological members is normalized from the division by the corresponding score of the initial 800-member ensemble, placing each component on the same scale.
- All scores except the MDCV function are oriented for minimization. However, the IGNS has the peculiarity of having negative values, making it necessary to establish a threshold (z_1) in the normalization. Thus, we establish $z_1 = -2$, since the preliminary analysis of selection under different scenarios (different catchments and number of members to be selected) showed minimum values for this score of about -1.5. With regard to the MDCV function, as testing different scenarios showed maximum values of about 0.8, we used a threshold of $z_2 = 1$. The hypothesis under the maximization of the MDCV is that a gain in dispersion should increase the reliability of the HEPS.

These six functions have been chosen because they quantify different aspects of ensemble prediction's quality. For instance, the CRPS simultaneously evaluates reliability, resolution, and uncertainty. The logarithmic or ignorance score assesses sharpness or spread and bias (strongly). Reliability is directly evaluated by the RD_{MSE}, while consistency and bias of the ensemble is assessed by the δ ratio. Finally, the maximization of the MDCV function (or minimization of the relationship z_2-MDCV) seeks to increase the spread of the ensemble.

For this study, we assumed a normal distribution for evaluating the CRPS and the IGNS. In these scores, we performed some simulations to estimate differences between empirical, Normal, and Gamma distributions. Results were omitted due to minor variations, in contrast to a high computational cost.

It is nonetheless important to note that this similarity is evaluated inside the ensembles with previsions varying between 30 and 800 hydrological members, as detailed below; in small samples, it is expected that the results represent the expected asymmetry of the information.

Regarding the reliability diagram, we evaluate conditional probabilities using an empirical distribution. With respect to rank histogram (or δ ratio evaluation), the equiprobable distribution of the HEPS members appears as a necessary condition. However, Anderson [15] emphasizes that, as a non-parametric method, the rank histogram (or Binned Probability Ensemble) does not depend on any of the details of the probability distribution of the forecasts or the initial conditions, so a large set of ensemble forecasts can be grouped for validation without difficulty. In the same way, Hamill and Colucci [76] established under a different hypothesis that:

> "each forecast should have independent and identically distributed (*iid*) errors. Nevertheless, it is also recognized that these are unrealistically ideal assumptions because any systematic error in the forecast model can result in forecasts with non-*iid* errors. Similarly, if the initial conditions are not equally plausible, but some are less likely than others, then the subsequent forecasts cannot be expected to exhibit equal accuracy".

Minimum number of members

With regard to the minimum number of members, which was arbitrarily defined as 30 here, the choice is mainly due to the high availability of initial members (800). For example, with 30 hydrological members, a level of simplification equivalent to 96.25 % of members is reached. It is certain that if the selection task had started with a pool of 50 members, then the minimum number of members could have been defined as 10, for example. Moreover, the minimum number of members is just a stopping criterion of selection with BGS because the number of members to define as optimal should focus on a specific analysis in each basin.

3.4.3 Combination of results

The variability of each experiment set-up in the cross-validation step increases the probability of reaching different hydrological member' selections. So, it is necessary to determine an integration mechanism for a global solution for each catchment. Here, the importance of each hydrological member $\mathbf{y}_i$ within the ensemble is then assumed as being directly proportional to the iteration number (iter) at which it was eliminated during the selection process in each of the five experiments (xp) proposed. The combined ranking is thus the mean rank of elimination given by:

$$\overline{R}\,(\mathbf{y}_i) \;=\; \frac{1}{5}\;\sum_{\mathrm{xp}=1}^{5}\mathrm{iter}_{\mathrm{xp}}^{\mathbf{y}_i}. \tag{3.2}$$

For example, if the rank of elimination of the hydrological member $\mathbf{y}_i$ is 50, 60, 200, 10, and 150 in the five experiments, then the mean rank of elimination is equal to 94. Finally, the final selection (s) of the nm[‡] best hydrological members corresponds to the hydrological members which have the highest mean rank of elimination given by:

$$s \;=\; \left\{\overline{R}_p,\; y_p\right\}_{p=1}^{nm},\quad \overline{R}_i \;\geq\; \overline{R}_j \;\text{ where }\; 1 \;\leq\; i \;\leq\; j \;\leq\; d. \tag{3.3}$$

It should be noted that another possibility to integrate the results could have been based on the frequency of selection of each hydrological member of the ensemble, and later to elect the members with the highest frequency, but as this integration leads to a low performance, this possibility was rejected.

[‡] nm is not necessarily equal to $nmin$ because nm reflects the analysis of the error on the validation set regarding the number of selected hydrological members.

3.4.4 Simplified HEPS - models participation

In the case of MEPS in which the members are not perfectly interchangeable (e.g. Meteorological Service of Canada (MSC), TIGGE database), the selection of hydrological members with BGS focuses directly on the combination of hydrological members that maintains or improves the characteristics of the super ensemble of reference.

But in the HEPS driven by a MEPS with interchangeable members (e.g. ECMWF EPS), the hydrological members selection should be directed more clearly to a method of weighting of hydrological models based on their participation in the final selected subset. Therefore, we can create a new simplified high-performance HEPS using the same proportion of the hydrological members associated with a random choice of the meteorological members.

For example, if the final selection shows that the simplified HEPS should consist of ten members for hydrological model "A" and thirty members for hydrological model "B", then we should expect to achieve a high performance HEPS if we randomly pick ten meteorological members to evaluate hydrological model "A" and thirty meteorological members to assess hydrological model "B". In Sect. 3.5.3 we present such an analysis.

3.4.5 Gain evaluation

Note that the CC could be used to compare the performance of the members' selection with respect to the 800-member set. So, in a general framework, if all features of the ensemble forecast have the same importance, one hydrological members' selection with equal performance to the 800-member set will lead to a CC equal to 5, values lower than 5 indicate a selection of higher performance than the base set of 800 members, and values greater than 5 indicate the detriment of any feature of the 800-member set. Hereafter this particular condition of unit weights in the CC will be called Normalized Sum (NS). This distinction is important to display the priority that can be defined a priori to any feature in the members' selection training with BGS. In this way, it is possible to define a gain index for the scores trade-off with respect to 5:

$$G_{\mathrm{NS}}(\%) \;=\; 100 \;\times\; \left(\frac{5}{\mathrm{NS}} \;-\; 1\right). \tag{3.4}$$

It is possible that the NS evaluated in the selected sets with BGS hides undesirable effects on the trade-off of the scores, for example to substantially improve one score with respect to the other ones. To check this condition, a gain index for each score is also proposed:

$$G_{\mathrm{SC}}(\%) \;=\; 100 \;\times\; \frac{\mathrm{Score}_{\mathrm{ie}} \;-\; \mathrm{Score}_{\mathrm{se}}}{|\mathrm{Score}_{\mathrm{ie}}|}, \tag{3.5}$$

where se and ie subscripts represent the selected ensemble of hydrological members and the initial 800-member ensemble respectively. A positive index indicates superior performance of the selected set. The absolute value in the denominator is needed to assess the performance of IGNS, which can have positive and negative values.

3.5 Results and analysis

Note that results discussed in this chapter correspond to a "pseudo test dataset" for comparing the performance between different scores in the process of selecting hydrological members, since the data used to minimize all error functions are exactly the same. It is a "pseudo test dataset" because there is a high probability that the data used in testing (the complete series) have been used in the BGS training process, becoming the indicator of an optimistic estimator of the selection [51]; however, we do emphasize that the first part of this research focuses on an analysis of scores in the BGS process with the subsequent integration of results.

Validation results were omitted mainly because they have a trend similar to the training ones, except for some experiments where the random distribution of the estimation and validation sets was not statistically homogeneous.

3.5.1 Selection performance

An example of the results obtained is shown in Fig. 3.4, which compares the 30-member and the 800-member results for the M04 catchment, after an optimization based on the δ criterion. In general the 30-member scores are better or as good as the reference set.

We stress the fact that the selection task focuses on the participation of the hydrological models. For instance, Fig. 3.4e shows that the selected hydrological members make use of 13 of the 16 available lumped models. However, the strong participation of models 3, 7, 9, and 14 is displayed, which is an interesting combination of hydrological models, especially taking into account the much poorer performance of the 16-member multi-model approach driven by the deterministic prediction (Table 3.3) and knowing that these hydrological models are not of equal quality with regards to MSE performance. This suggests that the selection favoured a diversity of errors.

Specifically, Fig. 3.4a shows that the 30-member CRPS is equal to the reference value corresponding to 800-member HEPS. Also, taking into account that the CRPS generalizes the MAE for a point forecast [66], it is important to stress that the CRPS values are always lower than the MAE values, when the deterministic counterpart was taken as the mean of each daily ensemble, in agreement with results obtained by other authors [21, 148]. Another remarkable feature of CRPS is its direct relationship with the streamflow magnitude; the shapes of the CRPS and of the hydrograph are similar. A direct strategy of optimization could then focus on removing the hydrological members that have a large impact on the daily extreme CRPS values. Note also that the selection not only preserves the mean CRPS (0.16), but also the structure of the CRPS series.

Figure 3.4b shows that the 30-member 4 % trimmed mean ignorance score (-1.01) has also improved over the initial value (-0.99). Regarding the time structure of the IGNS, it is

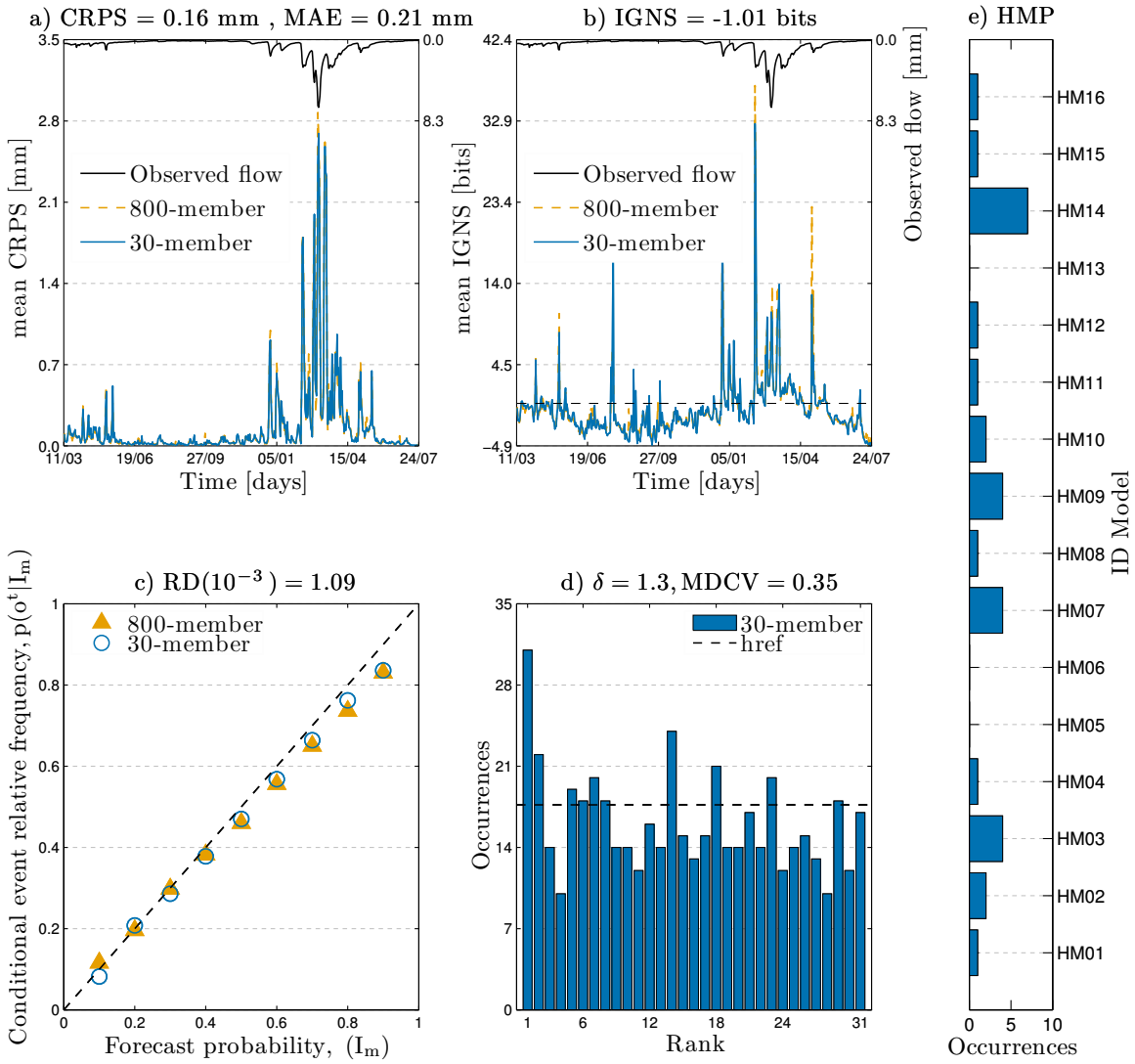


Figure 3.4: Comparison between the 800-member HEPS and 30-member HEPS for the 9th FTH for catchment M04. a) CRPS, b) IGNS, c) RD, d) RH, and e) HMP.

observed that both the 30-member and 800-member values have many extreme values which suggest low assessments of the predictive distribution of the ensembles, i.e. a bias problem in the forecasts (note that a value of 4.5 corresponds to an evaluation of the PDF near 0.0442).

With regard to the reliability diagram, Fig. 3.4c shows a considerable agreement improvement (1.09×10^{-3}) over the initial value (1.74×10^{-3}). This gain in reliability may be traced back to the optimization criterion used: the δ ratio, which is entirely based on the integration of the whole range in terms of corresponding verifications (observations). Similarly, Fig. 3.4d reveals that the rank histogram has a nearly uniform distribution, even if the first rank reflects a slight bias. Those imperfections demonstrate the difficulty inherent to minimizing the δ ratio.

At the end of the selection process, the MDCV has slightly decreased, from 0.37 to 0.35. This confirms that optimization with the δ criterion seeks diversity of the ensemble forecasts in the correct way, not necessarily maximizing the MDCV. Figure 3.4e illustrates the occurrence of each lumped model from the 30-member ensemble. A wide selection of models alone could justify the multi-model approach advocated here. Results show that 13 models out of 16 were selected in this case, and that no model was selected more than 7 times.

Taking into account the detailed analysis for the 30-member selections and the global analysis performed for each of the catchments, the CC leads to the best BGS results. The next section presents this analysis. However, the issue of the optimal number of hydrological members remains somehow blurred. So, Fig. 3.5 revisits that question in terms of the gain index based on the NS defined in Eq. 3.4. In this figure the vertical lines identify the iqr, the circles represent the median, and the diamonds correspond to the 10^{th} and 90^{th} percentiles.

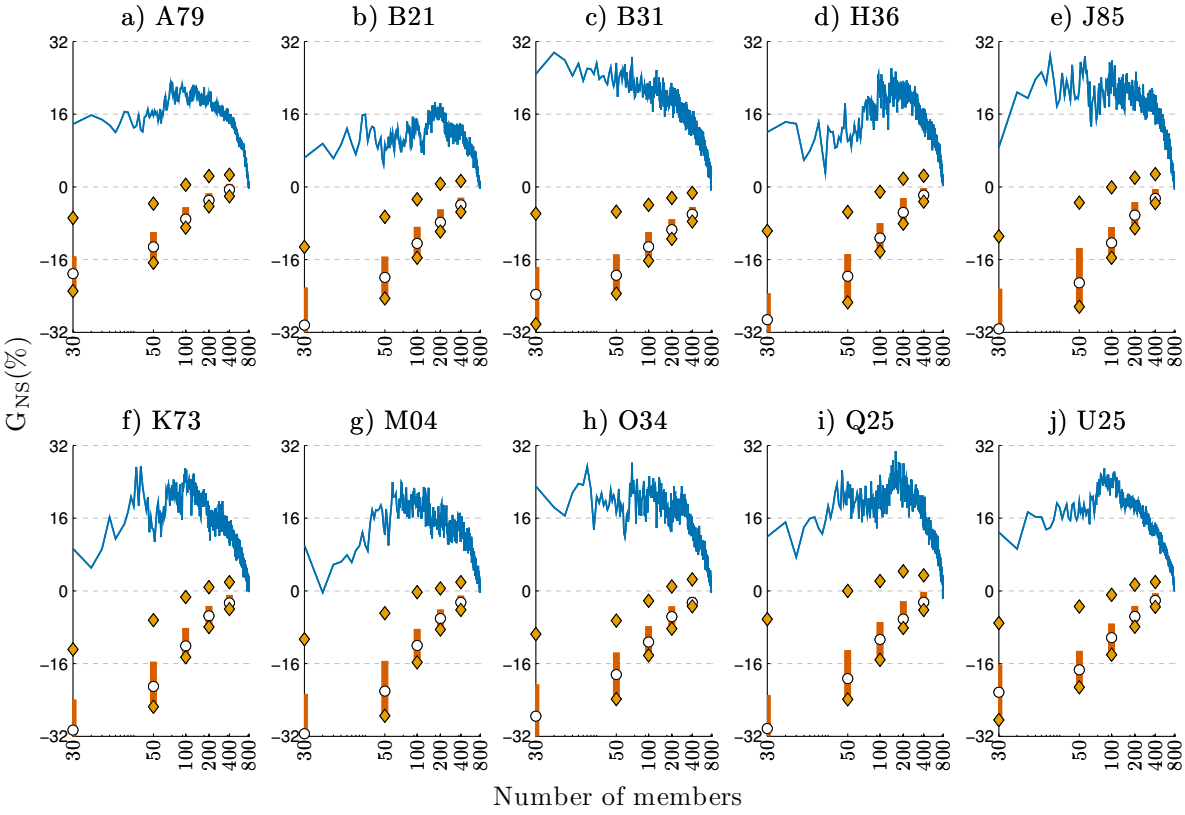


Figure 3.5: Evolution of the NS gain index for the 9^{th} FTH, optimization with the CC.

Figure 3.5 emphasizes that the 30-member selection always displays a positive gain index. However, one should keep in mind that the optimal number of hydrological members should be based on an individual analysis of the different scores trade-off, i.e. evaluating that the NS does not hide the detriment in a score(s) with gains made in other.

On the other hand, to reflect the BGS performance in the selection, Fig. 3.5 also presents the NS evaluation with 200 random selections of 30, 50, 100, 200, and 400 members in terms of gain index defined in Eq. 3.4. It is clear that BGS selection with positive gains are always obtained – improving the balance of the scores. Otherwise, in random experiments, the 10^{th}, 25^{th}, 50^{th}, 75^{th}, and 90^{th} percentiles are generally shown in the range of a negative gain index (i.e. a detriment to the balance of the criteria). This tendency is obviously stronger in random selections of 100 or fewer hydrological members where the probability of taking the most representative hydrological responses is lower. It is important to note how, even in the random selection of 200 and 400 members (25 % and 50 % of the 800 hydrological members), the NS in 75 % of the evaluations shows a negative gain index.

Table 3.5: Median of 200 random selections in catchment H36 for the 9^{th} FTH.

Members	CRPS	RD_{MSE}	δ	MDCV	IGNS	NS
30	1.01	1.50	1.80	1.05	1.11	6.47
50	1.01	1.25	1.53	1.03	1.06	5.88
100	1.00	1.09	1.28	1.02	1.02	5.41
200	1.00	1.02	1.11	1.01	1.01	5.15
400	1.00	0.98	1.03	1.00	1.00	5.01

To check each score individually, Table 3.5 shows the median of 200 random selections for basin H36 optimized with the CC. The random selections pick 50 hydrological members to evaluate each score in a standardized fashion, that is, dividing the score obtained in the selection subset by the reference score of all 800 members of base (see each component in Eq. 3.1 without weights). So, an analysis to evaluate the sensitivity of the scores with respect to the selection points out the following:

- In the hydrological members' selection, the greatest challenge is selecting a small set of members, for example 30 or 50.
- CRPS is indifferent to the selection of members, and to a lesser extent, both the low variability of the IGNS and the MDCV function.
- The hydrological members' selection presents its greatest challenges in maintaining or improving reliability and consistency of the ensemble represented by the δ ratio, as shown in Table 3.3. Therefore, to define the CC, such as an error term in BGS, the reliability term (RD_{MSE}) has more weight to guide the optimization in that direction. At this point, it should be noted that consistency has a direct relationship with reliability, although ensemble consistency does not necessarily imply that probability forecasts constructed from the ensemble are reliable in the sense of conditional outcome relative frequencies being equal to the forecast probabilities yielding a 45° calibration function on a reliability diagram, unless either the ensemble size is relatively large or the forecasts are reasonably skillful, or both [160].

Finally, Table 3.6 shows detailed results for each standardized score in the selection process with BGS for basin H36. It shows that the BGS methodology with the CC as error function, is not detrimental to any of the scores. Instead, gains in the balance scores (NS) are mainly due to the optimization of system reliability while preserving the quality of the other scores.

Table 3.6: Results of BGS in basin H36 for the 9th FTH with the CC.

Members	CRPS	RD_{MSE}	δ	MDCV	IGNS	NS
30	1.00	1.00	0.96	1.00	1.00	4.96
50	1.00	0.92	0.99	1.00	1.00	4.91
100	1.00	0.80	1.01	1.00	1.00	4.81
200	1.00	0.58	0.97	0.99	1.00	4.54
400	0.99	0.45	0.88	0.98	1.00	4.30

3.5.2 Scores interaction in the selection

Table 3.7 summarizes results for more catchments and optimization criteria. The 30-member comparison is based on a NS (Sect. 3.4.5). In this way, a value of NS lower than 5 indicates an improved performance. Performance for all criteria are also given for completeness and the best optimization criterion for each catchment is identified in bold letters.

Overall, the CC offers an effective and direct rule, finding balance between features offered by each of the criteria. However, it is important to point out the two cases for which the δ criterion provides a slightly better scores trade-off. This reflects the limitations of the BGS technique or the effects of the combination of results, because, if the objective function (CC) is equal to the criterion used to compare results obtained with different objectives, the CC should obviously always find the best solution within the vision of a global optimization tool.

The δ ratio criterion, based on a rank histogram which is the most common approach for evaluating whether a collection of ensemble forecasts for a scalar predictand satisfies the consistency condition [161], comes to a close second. It led to the best performance for two catchments and to the second best performance for five other catchments. This is particularly interesting considering the simplicity of this approach with respect to the combined approach. In addition, the δ criterion favoured the highest average participation of hydrological models.

The CRPS and IGNS led to a poorer selection, to the point that they were not considered further after experimenting with the first four catchments, allowing for an economy in computational time. The CRPS showed low variability, so it is not very sensitive to changes in the selection of hydrological members, as shown in Tables 3.5 and 3.6 previously. The IGNS demonstrated a negative relationship with reliability, leading to poor performance in terms of the RD_{MSE} and δ ratio. They are also correlated, optimizing one criterion often favouring the improvement of the other one.

Specifically the behaviour of the optimization of each score could also be described from the following relationships observed in Table 3.7:

- Optimization based on CRPS is detrimental to the reliability. For example, it increases RD_{MSE} by a factor of 10, for catchment Q25. The CRPS also decreases diversity of the hydrological members (MDCV), except for catchment B31 where it remained stable.
- The CC leads to stable CRPS values. The most remarkable gains come in terms of RD_{MSE}, as provided in the weights definition of Eq. 3.1. With reference to the δ ratio, evaluations reveal the difficulty in maintaining the stability of this criterion, but differences between the selection and the reference set are not pronounced. As for the MDCV, the diversity is, in most cases, maintained or improved. The IGNS performance is often slightly decreased. In conclusion, the CC promotes overall good performance, increasing the reliability of the system (decrease of the RD_{MSE} score) and ensuring the stability in the other scores.
- Selection based on the RD_{MSE} score is detrimental to the CRPS. As for reliability, there are some cases for which the error increases. This condition is surprising given that the CC always achieved reductions of this error, but this could not be attributed to the assumption of a greater weight of this score in the combination because the relationship is constant, which highlights the interaction between the scores as a mechanism implicit in the reduction of RD_{MSE}. The δ ratio is never improved, while diversity (MDCV) is lost except in three cases (B31, Q25, and U25) where interestingly the MDCV increased (theoretically consistent effect). Finally, the IGNS shows a negative trend with the minimization of the RD_{MSE}.
- By definition, the δ ratio focuses on the reliability and the consistency of the ensemble. In fact, it leads to better reliability performance in terms of RD_{MSE}, than when the selection is optimized with RD_{MSE} itself. The δ ratio also preserves the resolution of the forecast, as shown by the CRPS and IGNS results. All of this is accompanied by a slight loss in performance in terms of δ ratio, which can be explained by the direct relationship of this score with the number of members. However, this dependency rather than becoming an obstacle in the selection stands as a logical consequence of the system, since statistically a better performance is expected from a system that combines a larger number of members [7]. Finally, with respect to MDCV, it is shown once again that diversity, hypothetically represented by MDCV, fluctuates between values that indicate the extent to which such diversity needs to be maintained in the ensemble.
- When the selection process focuses on the maximization of MDCV, the relationship with CRPS, IGNS and δ ratio is always negative. However, there are four cases in which reliability is improved by increasing the MDCV, but while reliability improves, resolution drops.

In summary, the interaction of different scores, as seen from the 30-member selection, shows that the optimization focused on scores that mainly define the resolution of the ensemble (CRPS, IGNS) has a negative impact on the reliability, consistency, and ensemble diversity. It also reveals that, if the selection is based only on a reliability view, the ensemble loses resolution and consistency.

Table 3.7: 30-member HEPS scheme based on different scores for the 9th FTH.

Opt.cr	Basin	CRPS	RD_{MSE}	δ	MDCV	IGNS	NS	HMP	Basin	CRPS	RD_{MSE}	δ	MDCV	IGNS	NS	HMP
CRPS		0.24	7.0	4.3	0.34	0.41	5.7	8		0.21	4.0	4.5	0.49	−0.48	6.7	8
CC		**0.26**	**1.8**	**3.4**	**0.40**	**0.38**	**4.4**	**10**		**0.23**	**1.3**	**2.6**	**0.63**	**−0.16**	**4.7**	**13**
RD_{MSE}		0.26	2.8	4.6	0.40	0.49	5.0	7		0.23	2.5	3.9	0.53	−0.33	5.9	8
δ	A79	0.27	5.1	3.7	0.40	0.48	5.2	13	B21	0.23	2.1	3.0	0.56	−0.27	5.2	14
MDCV		0.28	11.1	5.0	0.46	0.65	6.8	7		0.24	5.2	3.7	0.61	−0.26	6.8	8
IGNS		0.24	9.6	4.8	0.31	0.38	6.5	6		0.22	23.2	8.0	0.39	−0.33	16.7	7
800-m.		0.26	5.0	3.3	0.41	0.44	5.0	16		0.23	2.4	2.2	0.57	−0.29	5.0	16
CRPS		0.18	5.9	4.6	0.22	−0.97	5.9	7		0.14	21.9	5.9	0.25	−0.96	17.2	6
CC		**0.13**	**0.9**	**2.0**	**0.23**	**−0.85**	**4.0**	**10**		0.16	0.7	1.8	0.37	−0.97	4.5	9
RD_{MSE}		0.15	3.5	5.2	0.24	−0.62	6.2	8		0.17	1.7	3.1	0.38	−0.84	6.0	5
δ	B31	0.13	2.9	3.3	0.23	−0.86	4.9	12	Q25	**0.16**	**0.6**	**1.6**	**0.37**	**−0.98**	**4.4**	**13**
MDCV		0.14	12.1	7.3	0.24	−0.70	8.7	7		0.18	3.9	3.5	0.45	−0.74	7.3	5
IGNS		0.12	17.4	7.1	0.17	−0.97	9.5	8		0.15	32.0	12.5	0.18	−0.41	26.9	6
800-m.		0.14	4.5	2.7	0.22	−0.88	5.0	16		0.16	2.2	1.5	0.37	−0.98	5.0	16
CC		**0.16**	**1.1**	**1.7**	**0.36**	**−0.97**	**4.5**	**11**		**0.16**	**0.5**	**2.3**	**0.39**	**−0.98**	**4.6**	**12**
RD_{MSE}		0.16	2.9	2.5	0.34	−1.00	5.5	7		0.17	2.3	3.1	0.35	−0.91	6.2	7
δ	H36	0.16	2.4	1.9	0.36	−1.02	4.9	13	J85	0.16	1.3	1.6	0.36	−0.99	4.6	13
MDCV		0.17	2.5	3.8	0.44	−0.79	6.4	6		0.18	1.6	2.5	0.44	−0.74	5.5	6
800-m.		0.16	3.5	1.5	0.37	−0.99	5.0	16		0.16	2.2	1.6	0.37	−0.98	5.0	16
CC		**0.16**	**1.3**	**2.4**	**0.36**	**−0.96**	**4.6**	**9**		0.16	0.7	1.9	0.36	−0.99	4.5	12
RD_{MSE}		0.17	3.4	3.7	0.35	−0.89	6.1	7		0.16	2.1	2.9	0.36	−0.92	6.3	6
δ	K73	0.16	2.1	3.3	0.33	−0.95	5.5	13	M04	**0.16**	**1.1**	**1.3**	**0.35**	**−1.01**	**4.4**	**13**
MDCV		0.17	2.5	4.2	0.43	−0.68	6.2	6		0.17	2.6	3.3	0.44	−0.75	7.0	5
800-m.		0.17	3.1	1.9	0.35	−0.93	5.0	16		0.16	1.7	1.5	0.37	−0.99	5.0	16
CC		**0.17**	**0.9**	**1.3**	**0.36**	**−0.87**	**4.1**	**13**		**0.29**	**1.2**	**2.9**	**0.37**	**−0.34**	**4.4**	**12**
RD_{MSE}		0.17	2.5	4.2	0.36	−0.67	6.6	5		0.30	2.9	5.0	0.37	−0.25	5.8	6
δ	O34	0.17	1.9	1.8	0.37	−0.85	4.7	12	U25	0.29	1.8	2.9	0.34	−0.32	4.7	15
MDCV		0.19	5.7	4.9	0.44	−0.51	7.9	4		0.30	3.0	3.7	0.43	−0.10	5.4	5
800-m.		0.17	3.5	1.5	0.36	−0.86	5.0	16		0.29	3.4	2.6	0.35	−0.36	5.0	16

Opt.cr and 800-m. represent the optimization criterion used in the BGS and the initial 800-member HEPS respectively.

Maximization of the MDCV is in general detrimental to the other criteria, but sometimes improves reliability, a condition that can easily be understood from a theoretical point of view. The δ ratio improves reliability while maintaining resolution. The combined approach stands out as the most balanced criterion.

The above analysis focused exclusively on 30-member selections. However, a global vision requires the analysis of the evolution of the scores as the number of hydrological members is reduced. Such an analysis is specific to each catchment. As an example, Fig. 3.6 shows evolution of the various scores as a function of the number of members for basin A79.

In order to assess the joint evolution of all scores, the gain index defined by Eq. 3.5 was used. Figure 3.6a and 3.6e clearly show that an optimization based on resolution of the system (CRPS or IGNS) is detrimental to the reliability. Figure 3.6 also highlights the correspondence of CRPS and IGNS throughout the selection process, when the optimization is focused on one or the other. RD_{MSE} optimization (Fig. 3.6b) is surprisingly unfavourable to the δ ratio (negative gain index), which is related to the indifference of the RD_{MSE} with respect to the

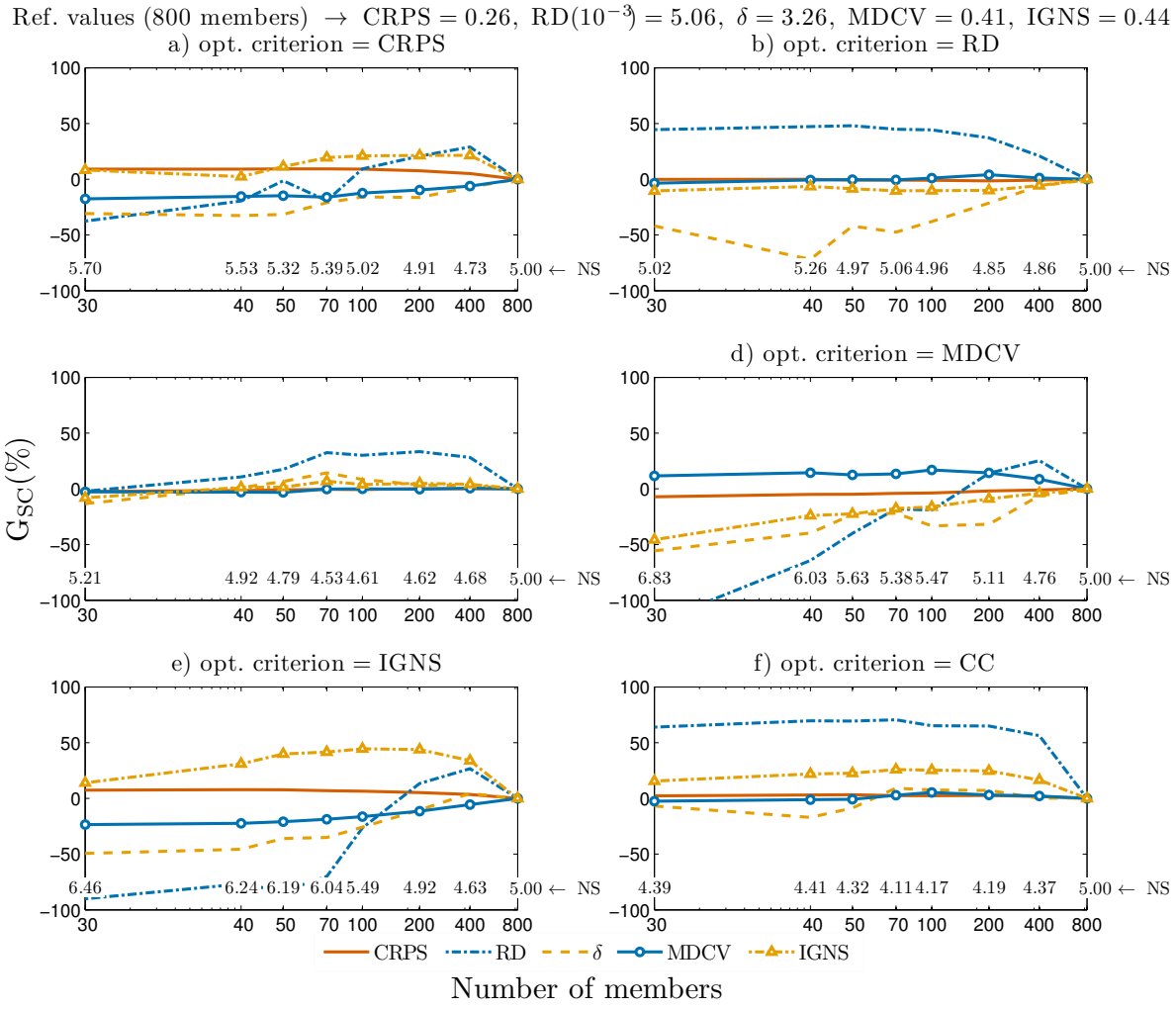


Figure 3.6: Evolution of the gain index for each score under different optimization criteria in the basin A79 for 9th FTH.

location of the observation within the ensemble, while this location analysis creates a solid indicator of the system consistency. Likewise, it is remarkable that the NS for RD_{MSE} is equal to 4.96 when the number of hydrological members is equal to 100. This is strictly because loss in consistency (negative gain index in the δ ratio of 40 %) and resolution (IGNS equivalent to losses of 10 %) is balanced by a positive gain of about 50 % in RD_{MSE}.

The δ ratio (Fig. 3.6c) displays a gradual overall improvement of individual scores in a selection of about 70 hydrological members, when the various scores show a tendency to decrease in performance. At this point it is important to note that the NS reached 4.53. Figure 3.6d shows that criteria focusing on resolution and consistency have a negative relationship with the maximization of the diversity (MDCV), overall gains are achieved only when the number of hydrological members is greater than 400. The CC (Fig. 3.6f) improves collective performance

of all scores in the selection, with an optimal number of hydrological members of 70 for this catchment, coinciding with the interaction shown in the minimization of the δ ratio (Fig. 3.6c). Scores tend to lose quality afterwards.

Table 3.8: Selection of 100 hydrological members based on the (CC) and δ ratio.

Opt.cr	Basin	CRPS	RD_{MSE}	δ	MDCV	IGNS	NS	HMP	Basin	CRPS	RD_{MSE}	δ	MDCV	IGNS	NS	HMP
CC		0.26	1.8	3.0	0.43	0.33	4.2	13		0.23	1.0	2.3	0.63	−0.19	4.4	14
δ	A79	0.27	3.5	3.0	0.41	0.43	4.6	16	B21	0.28	1.2	2.4	0.59	−0.28	4.5	16
800-m.		0.26	5.1	3.3	0.41	0.44	5.0	16		0.23	2.4	2.2	0.57	−0.29	5.0	16
CC		0.13	1.0	2.4	0.25	−0.83	4.2	14		0.16	0.4	1.3	0.40	−0.98	4.0	16
δ	B31	0.14	2.3	2.5	0.23	−0.85	4.5	16	Q25	0.16	0.6	1.4	0.36	−1.05	4.2	16
800-m.		0.14	4.5	2.7	0.22	−0.88	5.0	16		0.16	2.2	1.5	0.37	−0.98	5.0	16
CC		0.16	0.6	1.6	0.38	−1.03	4.2	14		0.16	0.4	1.5	0.39	−0.98	4.0	15
δ	H36	0.16	2.5	1.8	0.36	−1.04	4.8	16	J85	0.16	1.3	1.7	0.38	−1.00	4.6	16
800-m.		0.16	3.5	1.5	0.37	−0.99	5.0	16		0.16	2.2	1.6	0.37	−0.98	5.0	16
CC		0.16	0.6	1.7	0.39	−0.91	4.0	14		0.16	0.3	1.7	0.37	−1.00	4.2	15
δ	K73	0.16	2.6	2.2	0.34	−0.95	5.0	16	M04	0.16	0.8	1.3	0.36	−1.03	4.2	16
800-m.		0.17	3.1	1.9	0.35	−0.93	5.0	16		0.16	1.7	1.5	0.37	−0.99	5.0	16
CC		0.17	0.7	1.4	0.38	−0.87	4.1	16		0.29	0.9	2.2	0.39	−0.38	4.1	14
δ	O34	0.17	2.2	2.1	0.37	−0.89	4.9	16	U25	0.29	1.4	2.5	0.36	−0.42	4.3	16
800-m.		0.17	3.5	1.5	0.36	−0.86	5.0	16		0.29	3.4	2.6	0.35	−0.36	5.0	16

Opt.cr and 800-m. represent the optimization criterion used in the BGS and the initial 800-member HEPS respectively.

Table 3.8 groups the 100-member scores following optimization with the CC and the δ ratio, the two best ones. These values confirm the superiority of the CC, leading to the smallest NS for all catchments, mainly because of the great influence on minimizing reliability. This also maximizes MDCV to such an extent that it allows a proper balance between reliability, resolution, and consistency. It is also remarkable that for 8 catchments out of 10, the δ ratio is minimized even more than when the optimization is focused on the δ ratio itself. Optimization based on the δ ratio also improved scores over the initial 800-member values (NS$<$5) for 9 catchments out of 10. This single criterion is also very appealing, especially because it makes use of all 16 models in its selection. Additionally, the δ ratio can be highlighted as a simple optimization criterion, which for 100 % of the catchments makes use of the participation of all hydrological models in the formation of the solution, which is not the case for the optimization with the CC.

3.5.3 Interchangeability of MEPS members as input to hydrological models

In order to illustrate the interchangeability of the members of the ECMWF EPS and equiprobability of this system, Fig. 3.7a shows that a random selection oriented only with the HMP in the BGS has a chance to have even better performance than the 800-member HEPS upper 90 % (top of the box diagram). These box plots are constructed by retaining the participation

of hydrological models in the response but with a random selection of members of the MEPS. On the other hand, Fig. 3.7b shows the same kind of results under different random selections but without considering the participation of hydrological models found with BGS.

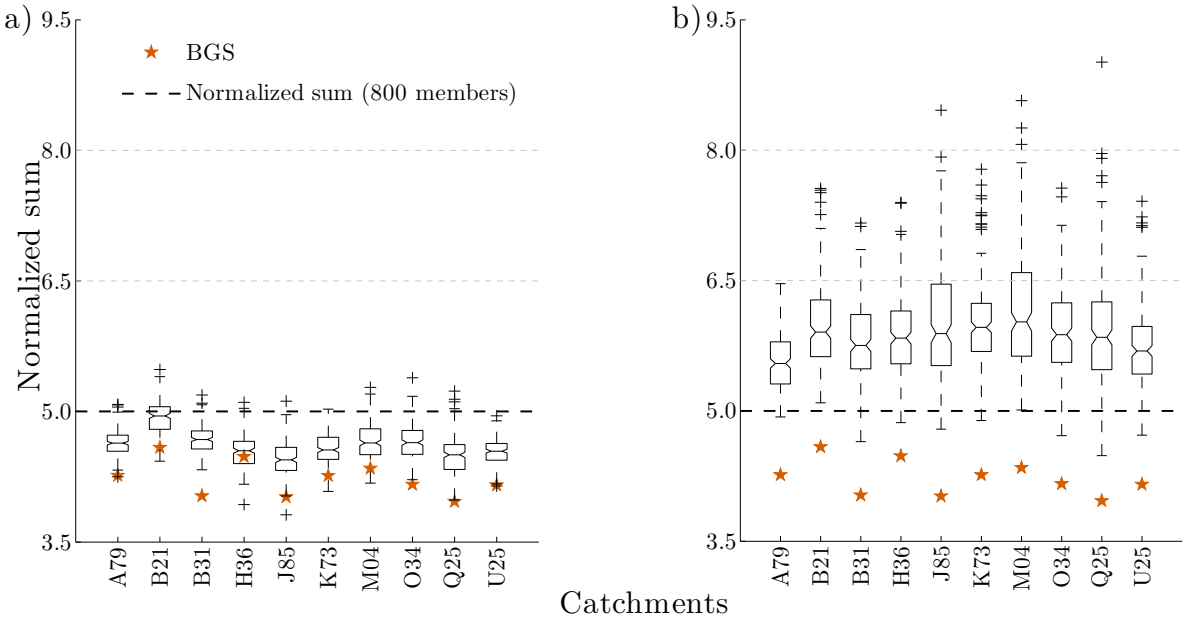


Figure 3.7: BGS and box-plots in 200 random experiments of 50 hydrological members for the 9th FTH. a) Random selection oriented with the frequency observed in the BGS to check the interchangeability in the 800 member-set; b) Random selection without any guidance to check the BGS performance.

Figure 3.7 highlights three main aspects: high-performance solutions based on the proportion given by the BGS, low variability, and high performance of the BGS solutions. The performance of selections based on the proportion of members found in the BGS solution is evident in Fig. 3.7a. So, it is demonstrated that the proportion of members for a hydrological model is generally a sufficient criterion to reduce the number of members while improving the balance of the scores represented by the NS. For comparison, Fig. 3.7b illustrates the system response to random selections without any a priori guidance, showing that in all cases the NS is greater than 5 and have recurring extremes greater than 7.

Regarding the variability of the NS evaluated in random selections guided by the BGS solution, it can be seen that the interquartile range ($Q_3 - Q_1$) is at worst equal to 0.3 (catchment H36), which is a much lower value than for the purely random selection, as shown in Fig. 3.7b where the latter interquartile range is equal to 0.6.

The generalization of the BGS method is discussed in detail in the next chapter, where the temporal and spatial generalization is evaluated for a nearby catchment. However, Fig. 3.7a shows that catchments H36 and J85 obtained combinations with a NS lower than those ob-

tained with the BGS method (see only cross points at the bottom in Fig. 3.7a), which can be associated with the integration of experiments carried out in a subdivision database for each catchment or the BGS algorithm structure – it is known that the classical BGS algorithm is unable to detect the collective influence of the variables.

3.6 Conclusion

Previous results on the number of hydrological members and the HEPS conformation [148] have shown, based on the database of the present chapter, that the ensemble predictions produced by a combination of several hydrological model structures and meteorological ensembles (800-member HEPS) have higher skill and reliability than ensemble predictions given either by a single hydrological model fed by weather ensemble predictions (50-member HEPS) or by several hydrological models driven by a deterministic meteorological forecast (16-member HEPS). So, our goal was focused on at least replicating the good quality of the 800-member set with fewer hydrological members.

Hydrological member selection is justified by the computational cost to issue a hydrological forecast based on the combination of meteorological models and hydrological models. In this line, the selection of hydrological members without sacrificing the quality of a forecast stands out as an operational option. Results presented here support the idea that selecting HEPS members is viable. It is, in general, even possible to expect a better balance of scores in the subset of selected hydrological members than in the much larger original ensemble, based on standard scores such as CRPS, IGNS, reliability diagram, and δ ratio. The diversity, sought in the multi-model approach with MEPS, may also be maintained in the final selection.

The simplification of the HEPS can be addressed from two points of view: as a function of the maximum simplification of the number of hydrological members or as a function of the maximization of the balance of the scores. Simplification of the number of hydrological members involves the definition of a limit ensuring statistical consistency of the scores assessed. A trade-off exists between the number of hydrological members and the level of improvement in scores. For example, in this study, the best balance of scores is achieved with a number of members fluctuating between 30 and 100, maximizing the qualities of the system: reliability, consistency, resolution, and diversity. So, in the worst case, this corresponds to an 87.5 % compression (700 members/800 members). The ultimate compression is in fact a compromise between the gain index and the complexity of the system. The ultimate decision should be established according to the requirements and operational capacity of the hydrological probabilistic forecast system.

The evaluation of six individual functions as criteria for optimizing the selection process revealed the complexity of the relationship between them. In many situations, improving one score is achieved at the expense of another score. Therefore, the design of a CC led to an im-

portant methodological improvement that integrates many characteristics of each score. The δ ratio is the best single optimization criterion, not very distant to the achievements of the CC. The CRPS is often the primary score used for evaluating HEPS performances. However, results here indicate that it is not a good choice for hydrological members' selection in this case of study. In fact, it was often possible to preserve or minimize the CRPS using other objective criteria. Likewise, the centralization of the selection process in the IGNS heavily penalized the reliability and consistency of the system.

With respect to the MDCV, the uncontrolled maximization of this parameter, which describes diversity, leads to a deterioration of the other sought qualities of the system. There exists a threshold beyond which the system abruptly loses reliability, resolution, and consistency. On the other hand, experiments showed that both the δ ratio and CC improve the balance of the scores.

The proposed methodology is part of the so-called data-driven models, so the design is independent of the database: in this case, the evolution of MEPS or hydrological models. This point stands out as one of the advantages of the proposed methodology, since the selection of hydrological members could be implemented in any desired combination between any MEPS (e.g. ECMWF EPS, MSC, US National Centers for Environmental Prediction (NCEP)) and hydrological models.

The cross-validation, a vital part of the proposed methodology, systematically deals with the issue of the short length of the series. However, it is widely applicable to any length series.

Finally, the encouraging results of this study will lead to an interest in testing other global search (non-greedy) tools such as EA (see Chap. 5).

Chapter 4

Generalization in Time and Space

In the previous chapter, we showed the efficiency of the simplification scheme, joining Cross-Validation (CV), Backward Greedy Selection (BGS), and the proposed Combined Criterion (CC). In this chapter, we assess the generalization ability of this scheme both with other Forecast Time Horizons (FTHs) and neighbouring basins.

That is, tests are made at two levels. At the local level, the transferability of the 9^{th} FTH hydrological member selection for the other 8 FTHs exhibits a 82 % success rate. The other evaluation is made at the regional or cluster level, the transferability from one catchment to another from within a cluster of watersheds also leads to a good performance (85 % success rate), especially for FTHs over the 3^{rd} FTH and when the basins that formed the cluster presented themselves a good performance on an individual basis. Diversity, defined as the hydrological model complementarity addressing different aspects of a forecast, was identified as the critical factor for proper selection applications.

4.1 Generalization test methodology

Figure 4.1 shows the generalization or test methodology of the hydrological members' selection at two levels: the local focuses on the extrapolation of results to different FTHs within the same catchment, while the regional level tests the temporal and spatial performance in nearby catchments, or under a broader perspective for the integration of regional results.

4.1.1 Extrapolation to different forecast time horizons

The HMP is performed on the results of sixteen hydrological models fed with the 9^{th} FTH of the ECMWF EPS. Thus, the application of this selection of hydrological members for the other eight FTHs (1 to 8 days) is a first level test. It has to be stressed that the idea of simplifying the Hydrological Ensemble Prediction System (HEPS) is only valuable if the HMP is invariant with regard to the FTH. However, one may always argue that the assumption of statistical

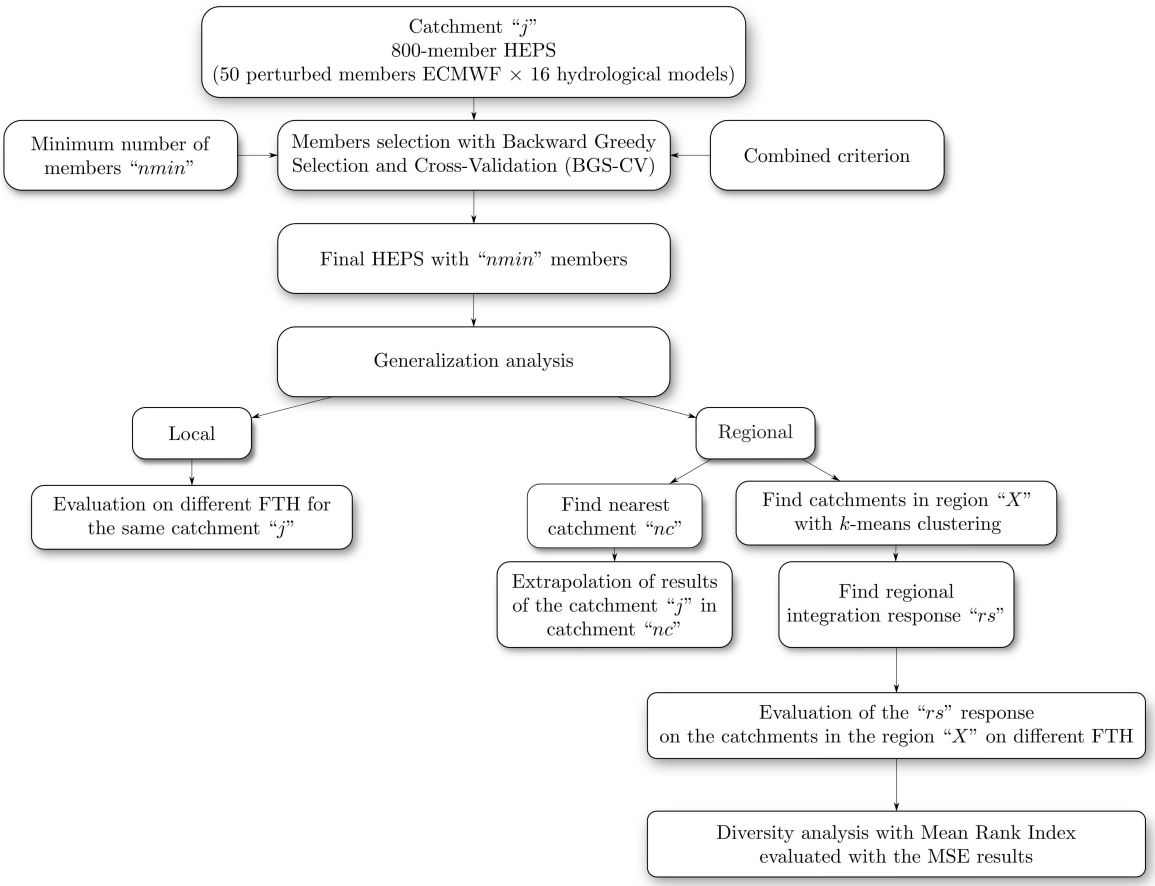


Figure 4.1: Generalization test for the evaluation of the HMP with BGS-CV.

independence between the test and training data, principally for FTHs closer to the ninth, may be somewhat questionable.

4.1.2 Extrapolation to a different catchment

Transferring HMP to a neighbouring catchment, and even further to a different FTH, constitutes a rigorous test of the generalization ability of results at both the temporal and spatial scales. The choice of the second catchment could first be viewed as a simple nearest neighbour problem. However, we explored the possibility of regionalizing the selection of hydrological members from the grouping of catchments by k-means clustering and subsequent integration of results to select the most representative hydrological members.

The k-means clustering algorithm (Sect. 2.2.1) is used to define 5 regions based on the combination of different characteristics of the catchments, such as geographic location of the basin outlet, minimum, mean, and maximum precipitation, evapotranspiration and streamflow (see Table 3.2). Every possible combination of features will yield a different distribution of catch-

ments that will be evaluated through the integration mechanism that will be presented in Sect. 4.1.2. Figure 3.1 shows an example of k-means clustering results based only on the geographic location of the basin outlets. Colours represent different clusters.

Regional integration mechanism

The results integration for region X, consisting of C catchments, is defined from matrix $\mathbf{S}$, which has C columns with $nmin$ rows representing the most $nmin$ important hydrological members as assessed by the mean rank of elimination ($\overline{R}$) for each catchment. Then, the process of forming a regional solution **rs** with q members is based on taking the most important members of each catchment without replacement until the number of members in **rs** is equal to the desired q, i.e. each member cannot be selected again later. Algorithm 2 details this procedure.

Algorithm 2 Regional integration mechanism pseudo-code

1. Determine the C catchments in the X region (clustering process).
2. Define the matrix $\mathbf{S} = \{\mathbf{s_1},\ \mathbf{s_2},\ \cdots,\ \mathbf{s_C}\}$
3. Establish the number of hydrological members q in the regional solution **rs**
4. Initialize $\mathbf{rs} = \{\}$, $h = 0$ and $i = 1$

repeat
  for $j = 1,\ \ldots,\ C$ **do**
    if $S_{i,j} \notin \mathbf{rs}$ **then**
      $\mathbf{rs} = \mathbf{rs} + S_{i,j}$
      $h = h + 1$
    end if
  end for
  $i = i + 1$
until $h > q$

Diversity evaluation

The participation of hydrological models in the regional selection stresses the importance of the integration of models with different characteristics. To view this in a deterministic framework, an index based on the performance rank assigned to each model in each catchment is proposed. Its calculation is summarized as follows:

- MSE for catchment i and hydrological model j is first calculated ($\text{MSE}_{i,j}$).
- Performances are next ranked for each catchment, leading to $\text{PR}_{i,j}$, for which the model with the lowest MSE is assigned the rank $\text{PR} = 16$ and the highest MSE is assigned the rank $\text{PR} = 1$.

- Finally, the mean rank of performance or rank index RI_j for each model is estimated based on the results of all 28 basins:

$$\mathrm{RI}_j \;=\; \frac{1}{28}\sum_{i=1}^{28}\mathrm{PR}_{i,j}. \tag{4.1}$$

4.2 Results and discussion

4.2.1 Selection process

The optimal number of hydrological members simplifying the HEPS was identified in the previous chapter to be between 50 and 100, depending on the catchment. In most cases, a significant gain with respect to the balance of the different criteria evaluated from the initial 800-member HEPS was achieved. Results presented in this section are based on a selection of 50 hydrological members.

Table 4.1 presents the results of the 50-member selection based on the CC, for 16 catchments uniformly distributed over France (see Fig. 3.1). The overall performance is the NS given by Eq. 3.1 with unit weights, values lower than 5 indicate a selection of higher performance than the base set of 800 hydrological members, and values greater than 5 indicate the detriment of any feature of the 800-member set. Beside each score is presented the gain index evaluated by Eq. 3.5. $\mathrm{RD}_{\mathrm{MSE}}$ values are expressed on a 10^{-3} basis.

To facilitate the visualization of results, Table 4.1 shows the performance of one selection oriented with the hydrological members' proportion found in the BGS-CV process. However, Fig. 4.2 and 4.4 present an analysis that shows the performance of multiple selections oriented by the BGS-CV solution and a random choice of the meteorological members from ECMWF EPS.

Table 4.1 shows that, in all cases, the NS is always lower than 5, indicating the superiority of the 50-member HEPS, even after a size reduction equivalent to a 94% compression of the initial 800-member HEPS (i.e. 750 members are removed).

Based on the gain score formulation (Eq. 3.5), it is noted that, for the 50-member selection, the CRPS and MDCV show low variability with mean gain indexes around 2 % and 5 %, respectively.

$\mathrm{RD}_{\mathrm{MSE}}$ shows a minimum gain of 49 % (catchment B21) and a maximum gain of 87 % (catchment K17), reflecting the emphasis given to this property in the formulation of the CC used in the selection process. With respect to the IGNS, index gains between −5 % and 27 % (excluding catchment B21) reflect an acceptable behaviour.

Finally, the δ ratio is the score more difficult to minimise or preserve; a positive index gain was obtained for only 25% of the cases (4/16), while the spread ranged from −39 % for catchment

Table 4.1: Selection of 50 hydrological members based on CC and BGS-CV process on the 9th FTH.

HEPS	Basin	CRPS	RD_{MSE}	δ	MDCV	IGNS	NS	Basin	CRPS	RD_{MSE}	δ	MDCV	IGNS	NS
50-m.	A69	0.284	1.3	1.5	0.67	0.39	4.0	A79	0.254	1.5	3.6	0.34	0.41	4.4
800-m.		0.284	7.0	1.8	0.78	0.37	5.0		0.263	5.1	3.3	0.44	0.41	5.0
Gain(%)		0	81	18	14	5			3	69	−11	23	−1	
50-m.	A92	0.183	0.3	2.3	−0.42	0.57	4.4	B21	0.232	1.2	2.6	−0.18	0.63	4.6
800-m.		0.192	2.4	1.8	−0.33	0.57	5.0		0.230	2.4	2.2	−0.29	0.57	5.0
Gain(%)		4	86	−28	27	0	-		−1	49	−16	−38	9	-
50-m.	B31	0.134	1.3	2.0	−0.84	0.24	4.0	H36	0.157	0.7	2.0	−1.02	0.36	4.5
800-m.		0.135	4.5	2.7	−0.88	0.22	5.0		0.161	3.5	1.5	−0.99	0.37	5.0
Gain(%)		1	72	27	−5	7	-		2	80	−37	2	−1	-
50-m.	H53	0.165	1.9	4.3	−0.76	0.36	4.6	H24	0.180	2.2	3.8	−0.82	0.37	4.6
800-m.		0.171	7.4	3.1	−0.71	0.33	5.0		0.185	7.1	2.9	−0.76	0.35	5.0
Gain(%)		3	74	−39	8	8	-		2	68	−32	9	6	-
50-m.	K17	0.205	0.5	1.8	−0.73	0.38	4.2	U25	0.290	0.9	2.6	−0.40	0.38	4.2
800-m.		0.213	3.6	1.7	−0.65	0.39	5.0		0.289	3.4	2.5	−0.36	0.35	5.0
Gain(%)		4	87	−9	12	−2			0	74	−1	13	7	
50-m.	J85	0.159	0.4	1.7	−1.00	0.40	4.2	K73	0.160	0.9	2.1	−0.93	0.38	4.3
800-m.		0.163	2.2	1.7	−0.98	0.37	5.0		0.165	3.1	2.0	−0.93	0.35	5.0
Gain(%)		2	80	−5	2	8			3	70	−5	0	9	
50-m.	M04	0.158	0.6	1.6	−0.98	0.37	4.3	M06	0.153	0.3	1.6	−1.09	0.39	4.2
800-m.		0.160	1.7	1.6	−0.99	0.37	5.0		0.159	1.4	1.5	−1.03	0.38	5.0
Gain(%)		1	68	−2	−1	2			4	79	−4	6	1	
50-m.	O34	0.166	1.0	1.6	−0.91	0.37	4.2	Q25	0.159	0.6	1.1	−0.94	0.39	4.0
800-m.		0.169	3.5	1.6	−0.86	0.36	5.0		0.163	2.1	1.4	−0.98	0.37	5.0
Gain(%)		2	71	1	5	3			3	73	22	−5	4	

50-m. and 800-m. represent the selection of 50 hydrological members and the initial 800-member HEPS respectively.

H53 to 27 % for catchment B31. Note that the δ ratio has an inverse relationship with the number of members of the selection, so it directly follows the complexity in maintaining the value of the initial 800-member HEPS in the selection process. Nonetheless, it was shown in the previous chapter that the δ ratio is the best individual metric in the simplification task.

4.2.2 Generalization test

Local analysis

For operational convenience, it is fundamental that the HMP for the 9th FTH is also appropriate for the eight previous FTHs. A lack of transferability of the HMP would considerably reduce the actual level of achieved simplification.

Here, temporal transferability is first evaluated comparing the NS of the performance of the 50-member selection to the 800-member performance, whose NS equals 5 in all cases. It is then compared to the performance of 200 random combinations with 50 hydrological members, in order to evaluate if any good performance may only be attributable to chance. Results for some of the FTHs and sixteen basins are gathered in box-plot diagrams (Fig. 4.2), where the

performance is based on random experiments that are set-up following these guidelines:

- Experiments considering the participation of hydrological models: taking into account the participation of hydrological models to assign to each model a number of members chosen randomly from ECMWF EPS.
- Without considering any "a priori" participation of hydrological models: hydrological members are picked randomly from the initial 800-member HEPS.

Results indicate that the median of 200 evaluations of 50-member oriented by the HMP is superior to the 800 reference members in 82 % of the evaluated cases. It is also noteworthy that, in only 11 % of the cases (14/128), this scheme leads to a worse performance than the 25 percentile of 200 random combinations test. Note that all these cases correspond to short FTHs (1 to 3 days), remarkably in the 2^{nd} FTH (Fig. 4.2a).

Another aspect that draws attention is the low dispersion of the BGS-CV selections represented by the interquartile range, highlighting the importance of the hydrological models participation in the selection process. Figure 4.2 also shows that the selection slowly loses efficiency as it moves away from the 9^{th} FTH. It also detects a systematic deficiency for catchment A69 and to a lesser extent for catchment B21. Nonetheless, these results are very encouraging.

Regional analysis

As described in Sect. 4.1.2, the regional analysis assesses the generalization ability of the HMP for a specific catchment with respect to another one. For example, Fig. 4.3 explores the transferability of the 50-member selection obtained for catchment Q25 for a 9^{th} FTH to catchment P72 for the 4^{th} FTH.

In general, Fig. 4.3 shows that results for the different scores are very similar for the 800-member and 50-member sets, except for the RD_{MSE} where the gain index reaches 51 %. In particular, Fig. 4.3a shows that the 50-member CRPS equals the reference value. Taking into account that the CRPS generalizes the MAE for a point forecast [66], it is important to stress that the CRPS values are always lower than the MAE values, when the deterministic counterpart was taken as the mean of each daily ensemble, in agreement with results obtained by other authors [21, 148].

Another remarkable feature of CRPS is its direct relationship with the streamflow magnitude; the shapes of the CRPS and hydrograph are similar. A direct strategy of optimization could then focus on removing the hydrological members that have a large impact on the daily extreme CRPS values. Note also that the selection not only preserves the mean CRPS (0.16) but also the structure of the CRPS series.

Figure 4.3b shows that the trimmed mean IGNS for the 50-member HEPS (−1.65) also presents an improvement over the initial value (−1.59). Regarding the time structure of the IGNS,

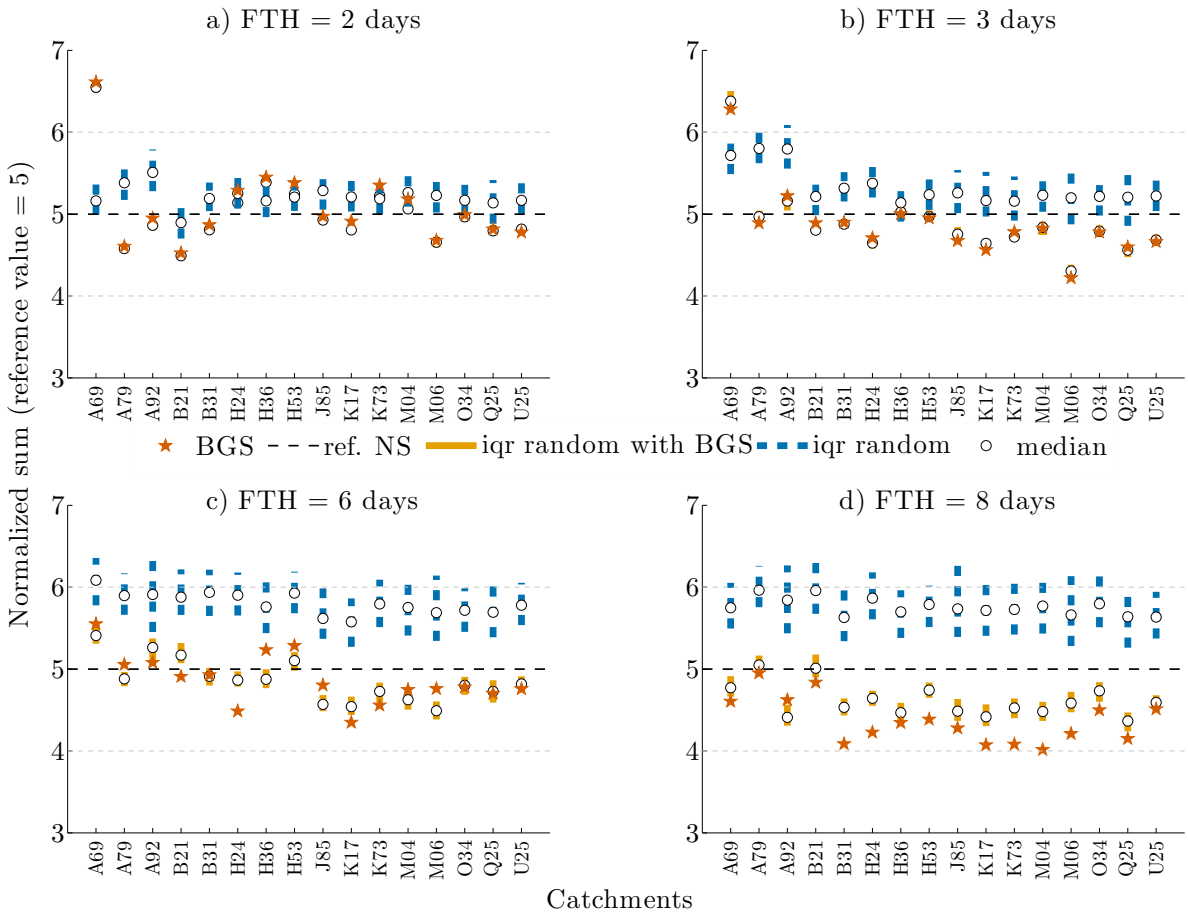


Figure 4.2: Evolution of the NS to evaluate the response sensibility with regard to the iqr of 200 random experiments in different FTH following these guidelines: (1) Considering the HMP found with BGS - CV (vertical solid bars), and (2) Random selection (vertical dashed bars).

it is observed that both the 50-member and 800-member series have high values for extreme events, showing a systemic problem in terms of ensemble bias.

With regard to the reliability diagram, Fig. 4.3c shows a considerable agreement improvement (4.21×10^{-3}) over the initial value (8.67×10^{-3}). This gain in reliability may be traced back to the optimization criterion used: the CC that focuses primarily on system reliability as defined by its weights. Similarly, Fig. 4.3d reveals that the rank histograms have a nearly uniform distribution, even if the first and the last rank reflect a slight bias. These imperfections demonstrate the difficulty inherent in minimizing the δ ratio.

Figure 4.3e illustrates the occurrence of each lumped model within the 50-member hydrological ensemble. A wide selection of models alone could justify the multi-model approach advocated here. Results show that 12 models out of 16 were selected in this case, and that no models were selected more than 9 times. Knowing that these models are not of equal quality with

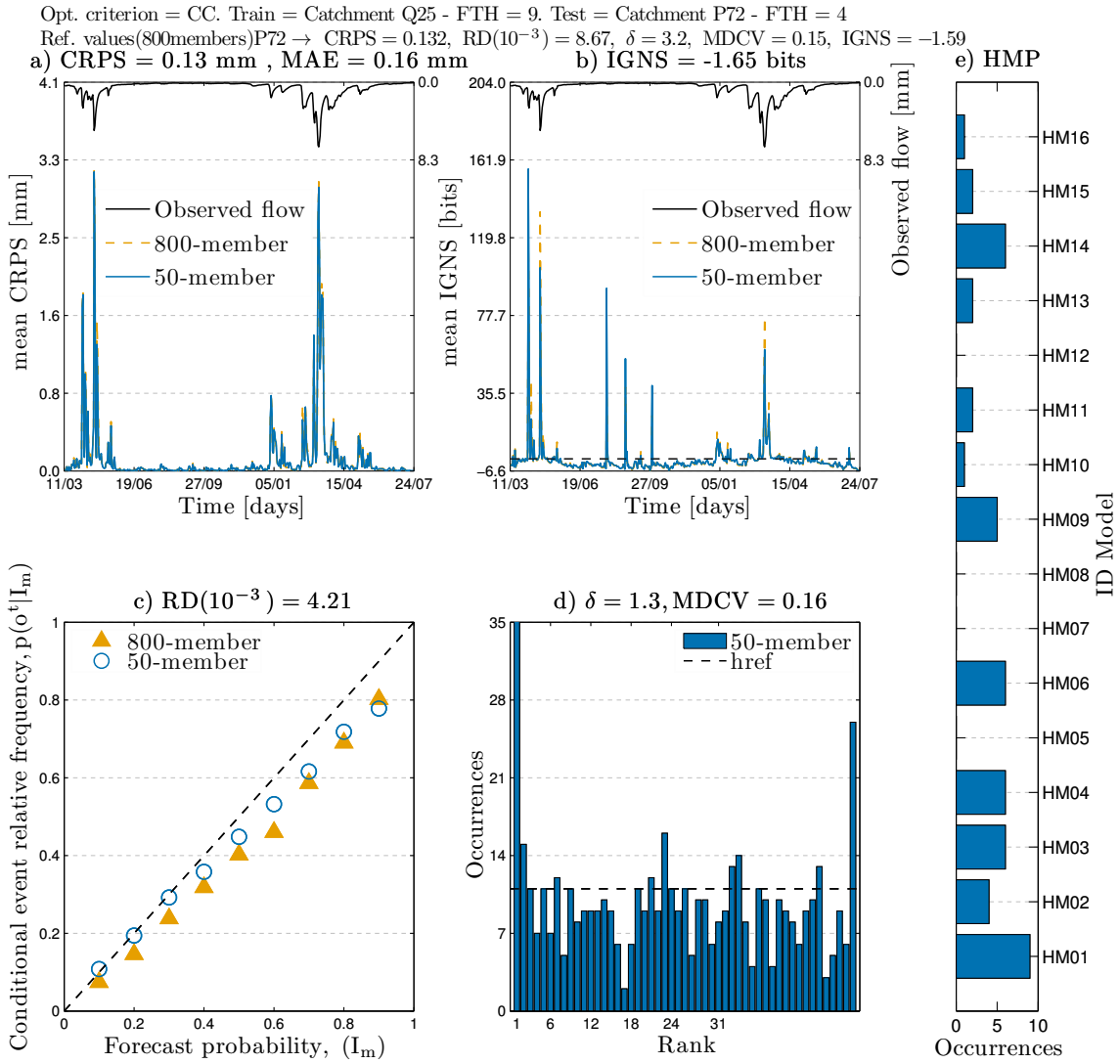


Figure 4.3: 800-member and 50-member HEPS comparison for the 9th FTH.

regards to MSE performance for instance, this suggests that the selection favoured a diversity of errors. At the end of the selection process, the MDCV had slightly increased, from 0.15 to 0.16.

To display an overview of the extrapolation of results to the nearest basin, Fig. 4.4 shows such an assessment under the same selection schemes analyzed in Fig. 4.2, i.e. analyzing various combinations considering or ignoring the solution found with BGS-CV. Each vertical bar represents the interquartile range (iqr) of 200 combinations of 50 hydrological members under the following guidelines: the combination is oriented with the HMP found with BGS-CV (solid vertical bars), the selection is completely random (dashed vertical bars).

Although, in general, the solution found with BGS-CV (stars in Fig. 4.4) exhibits the highest performance, given the interchangeability of MEPS members as input of hydrological models,

solutions focus on comparing the median of the evaluations that follow the HMP found with BGS-CV. Note the deficiency of the selections' extrapolation in basin A69 to basin A79, notably for early FTHs (2 to 5 days); these results do not appear in the figure because they are above 7.

Additionally, it is clear that the dispersion of the BGS-CV selections, evaluated from the interquartile range, is less than the one assessed in completely random selections. Likewise, the median of the BGS-CV selections is usually better than the reference set of 800 hydrological members, which corresponds to a NS equal to 5.

Another aspect that stands out in the extrapolation is the recurrent deficiency of selection in basins A69, A92, B21 and B31, i.e. 25 % of the basins tested. Initially, the deficiency in these basins at different FTHs shows the temporal consistency of HEPS, because if the deficiency of a given selection disappears at certain FTHs, it would reflect inconsistency of the selection task.

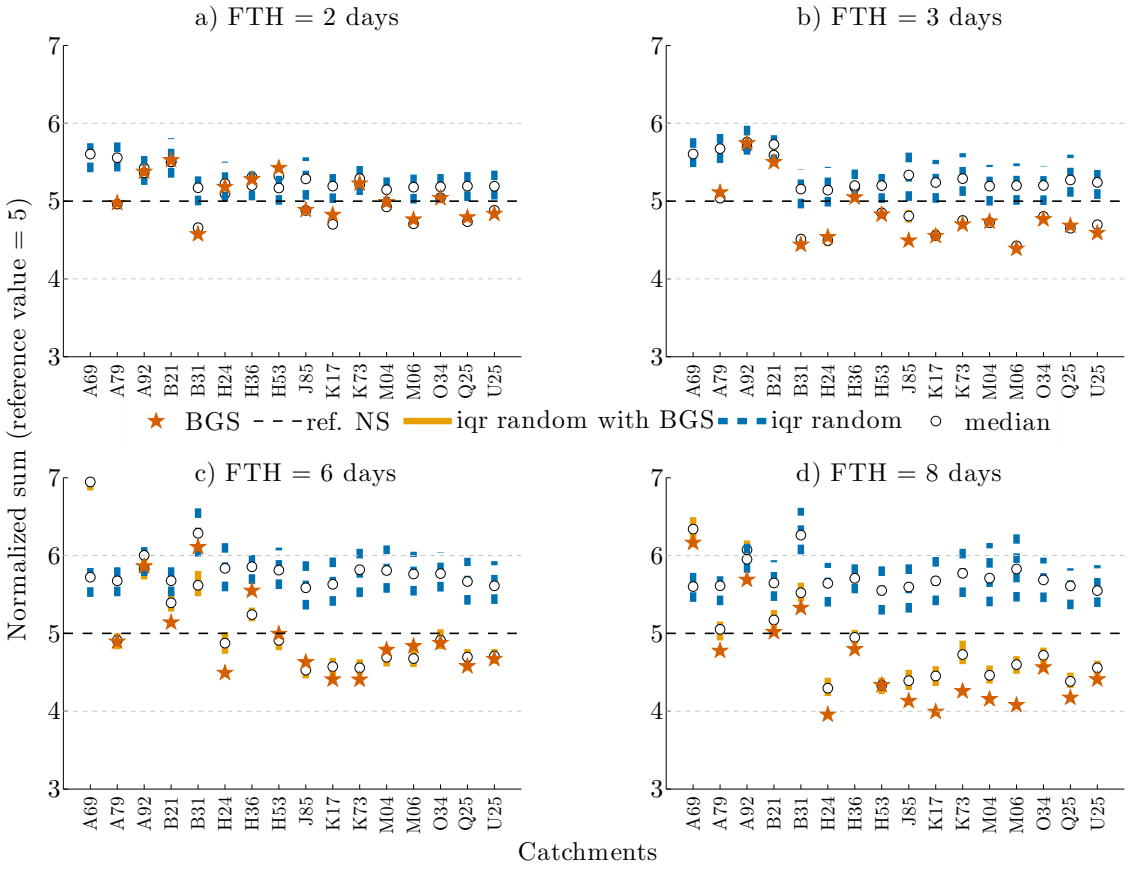


Figure 4.4: Evolution of the NS to evaluate the response sensibility of the extrapolation of results in the nearest catchments.

Likewise, it is noteworthy that extrapolation of the results of selection in basins A69, A79 and B21 are tested in basin A70; however, only the results of the hydrological members' selection

in basin A79 show considerable efficiency in most of the FTHs evaluated. It follows that while the geographic location of the basin outlet is an acceptable feature to run the extrapolation of results, it is not sufficient in some cases, requiring a more detailed analysis of other factors such as hydrometeorological and physiographic characterization of the basins.

The regional analysis that integrates several basins, which seeks to identify features that facilitate the combination of results, revealed that geographical location is the most important feature, followed by potential evapotranspiration, precipitation and streamflow, when the NS is used to evaluate the gain. However, consideration of the geographic location was found to be sufficient. Such results are presented in Table 4.2, after application of the k-means algorithm and the regional integration procedure already described in Sect. 4.1.2. See clusters distribution by colours in Fig. 3.1. Values lower than 5 determine that the scores of selection are better than the reference set. In each cluster, the catchments highlighted in bold represent the series that are not used in the evaluation of the HMP.

Table 4.2: Test based on the NS in new catchments and different FTHs of regional integration given by the analysis of clusters by geographical location of the basin outlets.

FTH	Cluster 1								Cluster 2						
	H24	K17	U25	**K13**	**K52**	**U06**	**U24**	**U27**	J85	K73	M04	M06	**H93**	**M15**	**M36**
1	5.08	5.25	5.06	5.19	5.36	5.20	5.15	5.12	4.96	5.19	5.09	5.07	5.06	5.09	4.96
2	5.17	5.18	5.12	5.07	5.24	5.02	5.36	5.04	5.03	4.97	4.97	4.89	4.85	4.90	5.00
3	4.89	4.85	4.87	4.71	5.01	4.60	4.86	4.78	4.66	4.63	4.67	4.73	4.71	4.70	4.67
4	4.50	4.56	4.69	4.26	4.76	4.53	4.68	4.59	4.67	4.57	4.72	4.71	4.70	4.71	4.60
5	4.82	4.56	4.56	4.31	4.85	4.54	4.76	4.68	4.70	4.33	4.51	4.54	4.40	4.43	4.29
6	4.99	4.74	4.86	4.59	4.87	4.59	4.76	4.79	4.41	4.47	4.53	4.29	4.49	4.53	4.34
7	4.50	4.52	4.42	4.58	4.74	4.50	4.52	4.50	5.01	5.04	5.00	4.81	4.77	4.80	4.80
8	4.38	4.25	4.27	4.16	4.71	4.22	4.33	4.33	4.43	4.61	4.78	4.62	4.47	4.84	4.41
9	4.50	3.97	4.09	4.04	4.36	4.07	4.32	4.17	4.09	4.32	4.59	4.39	4.31	4.39	4.22

FTH	Cluster 3				Cluster 4				Cluster 5				
	O34	Q25	**P70**	**P72**	B31	H36	H53	**H62**	A69	A79	A92	B21	**A70**
1	4.88	4.68	4.74	4.78	5.69	5.21	4.92	5.09	4.20	4.78	4.42	4.98	4.94
2	4.83	4.61	4.73	4.81	5.85	5.11	4.64	5.15	4.40	4.98	4.78	4.52	5.22
3	4.16	4.36	5.98	4.74	5.83	4.69	7.24	4.65	5.03	5.42	5.02	4.96	5.45
4	4.77	3.43	4.47	4.28	5.97	4.49	5.23	7.01	5.19	5.57	5.58	5.11	6.22
5	4.80	4.53	4.69	4.68	5.71	5.29	5.24	5.60	5.10	5.80	4.74	5.50	5.60
6	4.68	4.47	4.59	4.55	5.78	4.96	5.41	5.45	4.78	5.62	5.32	5.31	5.45
7	4.62	4.74	4.45	4.32	5.24	4.60	4.81	5.16	5.12	5.11	4.35	5.53	5.57
8	4.70	4.34	4.39	4.28	4.58	4.57	4.91	5.46	4.97	5.22	4.25	5.50	5.08
9	4.36	4.15	4.28	4.12	4.26	4.08	4.50	4.74	4.87	4.66	4.45	4.92	5.38

Note that results in Table 4.2 are due to the evaluation of one combination of MEPS members randomly chosen, but respecting the participation of hydrological models found with BGS-CV. Additionally, for purposes of extrapolation of results, in the evaluation of the NS, a threshold z_1 equal to -4 was used, because in the firsts FTHs (1 to 4 days) some values lower

than −2 were obtained for the trimmed mean IGNS. In Table 4.2, the NS for the 9th FTH is generally lower than 5 for catchments subjected to the regional integration (except basin A70). Furthermore, in 44 % of such assessments (catchments H24, K17, U25, J85, K73, H36, and H53), the regional integration presents better results than the local performance relative indicators shown in Table 4.1.

Although the regional integration in clusters 1, 2 and 3 shows that the 85 % of the NSs are lower than 5 and the remaining 15 % corresponds principally to the first FTHs (1 to 3 days), the clustering and posterior regional integration is less efficient for groups 4 and 5, whose NSs are higher than 5 in 65 % of the cases.

The behaviour in cluster 5 is inherited from the low extrapolation efficiency highlighted in basins A69, A92, and B21 (Fig. 4.4). As such, the proposed regional integration mechanism is shown as a consistent task, since its efficiency is a function of performance of its components.

With regard to cluster 4, the regional solution shows a lower diversity of hydrological models. This factor is evident in Fig. 4.5 which illustrates that, for this cluster, 70 % of the hydrological members originate from only three hydrological models (HM03, HM06, and HM14), which is quite a different behaviour than for clusters 1, 2 and 3 where the proportion of the three most selected models reaches 58 %, 56 %, and 44 %, respectively.

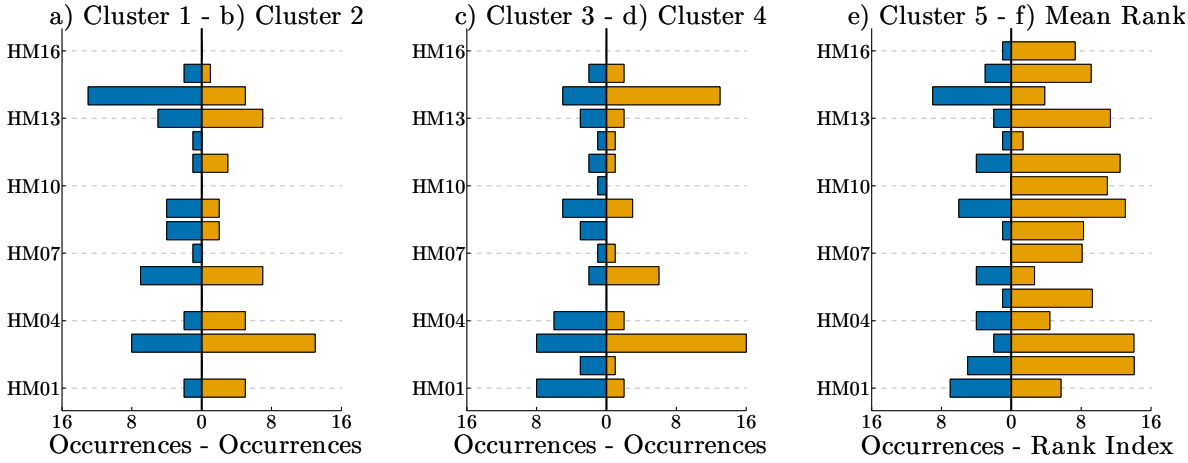


Figure 4.5: Hydrological Models Participation (HMP) in five clusters (from a to e) and mean models rank index evaluation (f).

Thus it seems that diversity, as a characteristic of the final selection of hydrological members, appears to be a factor with a significant impact on the performance of the selection. In other words, the participation of hydrological models in the regional selection stresses the importance of the integration of models with different characteristics. To view this in a deterministic framework, the index based on the performance rank assigned to each model in each catchment (Sect. 4.1.2) shows that the most selected models (HM01, HM03, HM06, HM09, and HM14)

occupy quite different ranks (Fig. 4.5). For instance, HM03 and HM09 present a high performance while HM01, HM06 and HM14 are of lower performance. This feature exemplifies the notion of diversity discussed in different scientific communities concerning ensemble methods.

Diversity can be defined as the search for models that complement their skills, so that each model focuses on different objects. Diversity in the ensemble is thus a vital requirement for successful modelling. In practice, it appeared to be difficult to define a single measure of diversity and even more difficult to relate that measure to the ensemble performance in a neat and expressive dependency [92]. Nevertheless, the regional clusters in Fig. 4.5 make use of most of the 16 available models, whatever their performance rank. For example, the most frequently selected models in cluster 2 are HM03 and HM06 despite the fact that HM02 exhibits the same rank of performance as HM03 and that HM06 presents one of the lowest ranks in the ensemble.

4.3 Conclusion

This chapter has focused on the generalization quality in time and space of a 50-member HEPS selected from the 800-member ensemble at the 9$^{\text{th}}$ FTH. When applied to the other 8 FTHs, the 50 selected members also improved performance over the initial 800-member HEPS in 82 % of the situations. It was particularly successful when applied to a nearby catchment of the same cluster. Member diversity seems to be the key to this simplified HEPS that makes use of only 6.25 % of the initial structures (50 members/800 members). Indeed, it has been shown that most 50-member HEPS relied on a broad selection of hydrological models, which gives further support to the multi-model hydrological approach.

Comparing scores obtained for the 50 representative hydrological members to the ones of the initial 800-member ensemble indicated that the proposed selection methodology, which is based on cross-validation and the combination of scores into a single function, generally leads to a good performance in terms of gains of individual scores. However, these gains were not entirely transferable under the scheme of extrapolation evaluated here. This drawback may in part be attributable to the simple selection methodology used here along a linear integration of scores that has no real control over balance, or the need to evaluate more features to enhance such transferability in the clustering approach. Finally, results of this chapter encouraged us to explore the following guidelines in the next chapter:

- Optimize several performance diagnostics simultaneously or find a Pareto set of solutions identifying trade-offs among the various performance metrics.
- Combine the HMP with meteorological members chosen through a technique similar to the one proposed by Molteni et al. [109], instead of picking them randomly.

Chapter 5

Comparison of Techniques in a General Framework of Selection

In Chap. 3 and 4, we showed the efficiency of a simplified Hydrological Ensemble Prediction System (HEPS) based on the Hydrological Models Participation (HMP). Furthermore, we exposed the antagonism between bias, significantly represented by the IGNorance Score (IGNS), and the reliability, estimated directly with the Reliability Diagram (RD).

In this chapter, we compare the performance of various schemes to find out indirectly the "optimal" HMP. Thus, in a given catchment for the 9^{th} Forecast Time Horizon (FTH), the HMP is explored with four techniques: Backward Greedy Selection (BGS), Nondominated Sorting Genetic Algorithm II (NSGA-II), Linear Correlation Elimination (LCE), and Mutual Information (MI). The HMP will indicate the number of representative members to propagate into each hydrological model, while generalization is evaluated in a neighbouring catchment at different FTHs.

With the aim to highlight the importance of the HMP as a simplification base, the different simplification schemes are compared with the evaluation of an intuitive scheme of uniform HMP. In the latter, we evaluate the propagation of three representative members from the European Centre for Medium-range Weather Forecasts (ECMWF) - EPS into 16 hydrological models, which leads to a 48-member HEPS.

The results showed that the difficulties in simplifying mainly originate from the preservation of the system reliability. Compared with the efficiency shown by BGS and NSGA-II, both the uniform HMP scheme and simplification schemes based on members' correlation (LCE, MI), showed generally poor performances.

5.1 General framework of the simplification scheme

Following the assumptions made in Sect. 3.3, the simplification of the 800-member HEPS, resulting from the combination of 50 equiprobable scenarios of precipitation and 16 independent hydrological models, is based on the HMP.

In the two preceding chapters, we showed that in terms of the HMP for a given number of hydrological members, the propagation of certain meteorological members, chosen randomly through corresponding hydrological models was sufficient to achieve a simplified HEPS of at least the same performance as the 800-member HEPS. However this methodology left open some questions that will be addressed later in this chapter.

For example, the notion of representative precipitation members of the ECMWF - EPS is analyzed based on the evaluation of the k-means clustering technique. Also the performance of BGS and NSGA-II is compared with random selections and methods based solely on the correlation of the hydrological members.

5.2 Methodology for the simplification techniques comparison

In Fig. 5.1 we illustrate the scheme of simplification that serves as a model for comparing the four selection procedures evaluated here: BGS, NSGA-II, LCE, and MI. In this figure, we outline how to train and test systematic selection procedures. For training, the 800-member HEPS database of a given catchment is analyzed for the 9^{th} FTH. This FTH was chosen because it is generally the best scenario evaluated in reference to the 800-member HEPS (see Fig. 3.2). Later, selection techniques are used to infer the HMP in the simplified scheme.

To test the different methods, we apply the simplification scheme based on the HMP in a nearby catchment in all FTHs, in order to assess the HMP generalization ability in space and time. The application of the HMP as a simplification base is similar to that presented in Sect. 5.1. Additionally, we assessed random evaluations and uniform HMP approaches to establish a reference for selection complexity.

However, to assess hydrological models with representative precipitation members, as proposed by several authors [57, 83, 100, 109, 164] and following the trend of recent clustering information available at ECMWF - EPS [60, 118], the HMP directly orients the evaluation of representative precipitation members at each time step to subsequently propagate them into their respective hydrological model. For this, we rely on the k-means technique (see Sect. 2.2.1) configured with the Euclidean distance as the similarity measure, and the HMP to define the number of clusters, so the EPS member closest to its cluster centroid will define the representative member.

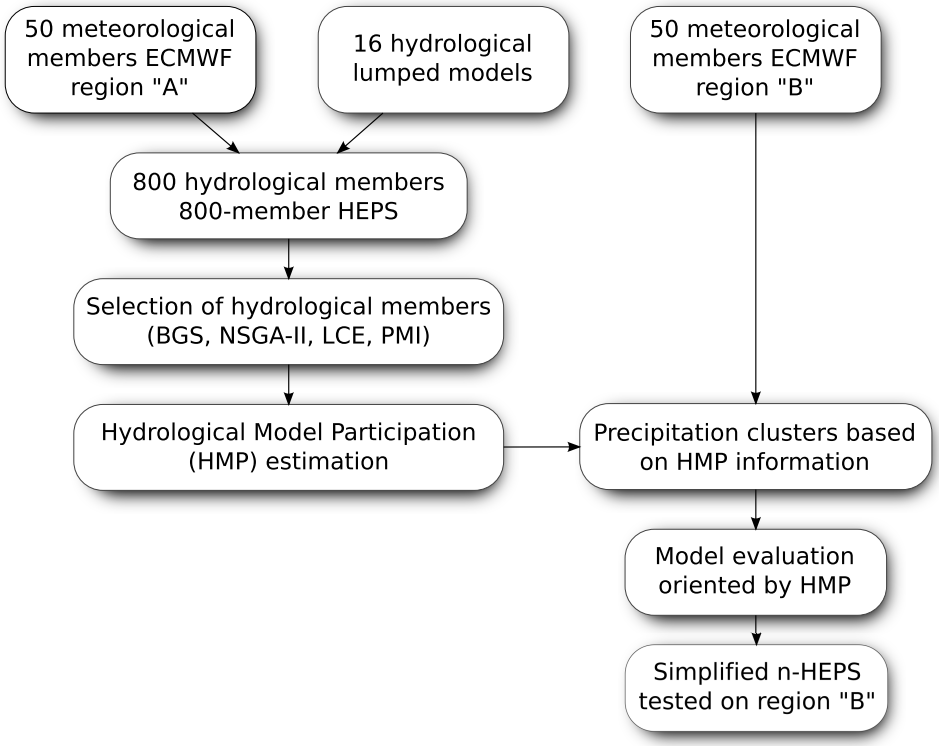


Figure 5.1: HEPS simplification based on ECMWF EPS clustering and different HMP schemes.

Consider the previous example (Table 3.4), initially it is sufficient only to determine the centres of seven clusters in the meteorological information at each time-step to propagate them into hydrological model HM#1, and so on, to propagate the ten representative precipitation members into hydrological model HM#16. Finally, we evaluate the simplified model quality taking as a reference the 800-member HEPS.

5.2.1 Probabilistic properties to evaluate

The analysis presented in Chap. 3 showed that the best score in the selection with BGS was the δ ratio, which effectively combines the bias, reliability, and ensembles consistency; however, its interpretability is difficult because it is proportional to the number of ensemble members. On the other hand, the CRPS showed low sensitivity in the simplification process. Thus, in this chapter, the selection focuses on the preservation of two scores that represent the duality of the bias and reliability in probabilistic forecasting: the Ignorance score (see Sect. 1.3.2) and the error in the reliability diagram (see Sect. 1.3.3).

The IGNS strongly penalizes the bias; our objective is minimization. For its evaluation, based on the ensemble PDF, we do not assume an a priori PDF (unlike the assumption of normality predefined in Chap. 3 and 4). Instead, at each time step, we evaluate the ensemble PDF for a given value y with a Gaussian-kernel estimator:

$$
\begin{aligned}
f\left(y,h\right) &= \frac{1}{dh}\sum_{i=1}^{d} K\left(\frac{y-y_i}{h}\right), \qquad (5.1)\\
\text{where:}\quad K(x) &= \frac{1}{\sqrt{2\pi}} e^{-x^2/2},\\
h &= \frac{min\left(0.9\ \sigma(\mathbf{y}^t), 0.75\ \text{iqr}(\mathbf{y}^t)\right)}{\sqrt{d}}.
\end{aligned}
$$

K represents the heights of all the kernel functions contributing to the smoothed estimate at a given value, y, and h is the bandwidth or smoothing parameter. It is evaluated, as suggested by Silverman [137], based on the standard deviation, $\sigma(\mathbf{y}^t)$, the interquartile range, $\text{iqr}(\mathbf{y}^t)$, and the number of data (members in this case), d. The optimal size of h is similar to that of the class width in a histogram. As in other methods of application of windows, it may cause under or oversmoothing [89].

With respect to the evaluation of the reliability, the objective is to match the conditional observed probability to the evaluated probability, so we calculate these differences directly into the diagram of reliability with the classic MSE (RD_{MSE}). This measure is negatively oriented, i.e. we seek to minimize it. To evaluate the conditional observed probability, it is necessary to establish the confidence limits; for this we use the same kernel function type shown in Eq. 5.1. Subsequently the observed frequency is evaluated as shown in Sect. 1.3.3.

5.2.2 Number of members in the simplified scheme

In the two preceding chapters, it was shown that, in general, between 30 and 70 hydrological members were sufficient to ensure a simplified HEPS with at least the same performance as the 800-member HEPS. Consequently, in this phase of selection technique comparison, we predefine 48 hydrological members in the simplification with the aim of performing comparisons with the two schemes that serve as reference. The first consists in a uniform participation of three forecasts for each of the sixteen hydrological models, resulting in a 48-member HEPS. In that case, the three representative members or ensemble cluster centres from the ECMWF - EPS are estimated and propagated into the sixteen hydrological models. The second scheme consists of a random choice of HMP limiting itself to 48 members, because it is clear that the selection sensitivity is directly related to the final number of hydrological members (see Tables 3.5 and 3.6).

5.2.3 Datasets

Results presented in this chapter are based on a selection of four basins that show different behaviours according to simplification. Basins H24, H93, M06, and P70 are used for training, whereas corresponding neighbouring basins H36, M04, M15, and P72 are used for testing.

Catchment location is illustrated in Fig. 3.1 and their main characteristics are given in Table 3.2. It is important to note again that the training is implemented only in the 9$^{\text{th}}$ FTH while the generalization test runs for all FTHs of the neighbouring basin, e.g. if we obtain a simplification scheme in basin H24 for the 9$^{\text{th}}$ FTH (training phase), its generalization ability will be evaluated in basin H36 for all FTHs, from first to ninth (testing phase).

5.2.4 Selection techniques setup

Regarding the optimization criterion, a combined approach with similar guidelines set out in Sect. 3.4.2 is proposed in Eq. 5.2 for LCE, MI, and BGS; however, in this case, we estimate a normalization threshold for the IGNS equal to -3, as well as an allocation of unit weights to each component.

$$\text{CC} \;=\; \frac{-3 \;-\; \overline{\text{IGNS}}_{\text{se}}}{-3 \;-\; \overline{\text{IGNS}}_{\text{ie}}} + \frac{\text{RD}_{\text{MSE}_{\text{se}}}}{\text{RD}_{\text{MSE}_{\text{ie}}}}, \tag{5.2}$$

where each score in the selected ensemble of hydrological members (se subscript) is normalized by the corresponding score in the initial 800-member ensemble (ie subscript), placing each component on a similar scale.

Note that manipulation weights offers the possibility of constructing trade-off among different objectives known as Pareto fronts; however, this formulation does not provide a necessary condition to find the optimal Pareto front, x^* , that is, solutions for which do not exist another points $x \in X$ such that $obf_i(x) < obf_i(x^*)$ for at least one function [99], here obf represents one of the objective functions, i.e. IGNS or RD_{MSE}. A more detailed description of the techniques presented below can be found in Sect. 2.4.

Linear correlation elimination

As discussed in Sect. 1.1.3, in the ensemble context, the manipulation of the negative correlation between predictors is related to the MSE reduction, therefore this property is the basis of some training methods of prediction ensembles such as ANN ensembles training based on negative correlation [37].

Accordingly, we propose a filter type selection based solely on the correlation between hydrological members without involving the probabilistic scores. With this objective, our approach based on the work of Haindl et al. [75], is shown in Algorithm 3, in which the less correlated pair of members is defined as the first two members of the selection, $\mathbf{s}$. So, a first set of candidate members is calculated as the relative complement of $\mathbf{s}$ in $\mathbf{Y}^*$. With each iteration we add the member that most decreases the average correlation. The latter is evaluated with respect to the selected members in a previous iteration.

*The relative complement of $\mathbf{s}$ in $\mathbf{Y}$ (also called the set-theoretic difference of $\mathbf{s}$ and $\mathbf{Y}$), denoted by $\mathbf{Y}\backslash\mathbf{s}$, is the set of all elements which are members of $\mathbf{Y}$ but not members of $\mathbf{s}$.

Algorithm 3 Sequential LCE.

1. Define the features space (members space):
   $\mathbf{Y} = \{y_1, y_2, \ldots, y_d\}$, where $d = 800$ in this case.
2. Define the number of members to select nm.
3. Calculate the correlation matrix of all features or members:
   $\text{corr}\,(\mathbf{y}_i, \mathbf{y}_j)$, for $i, j \in \{1, 2, \ldots, d\}$.
4. Evaluate the first two selected members as the two members less correlated:
   $$\{s_1, s_2\} = \underset{y_a, y_b \in \mathbf{Y}}{\text{argmin}}\ \text{corr}\,(y_a, y_b)\,.$$
5. Define initial candidate members set, $cm = \mathbf{Y} \backslash \mathbf{s}$, and initialize $iter = 3$

repeat
  for all $\mathbf{y}_k \in cm$ **do**
    6.1 Evaluate average correlation
    $$\overline{\text{corr}}_k = \frac{1}{iter - 1} \sum_{m=1}^{iter-1} \text{corr}\,(\mathbf{y}_k, s_m)$$
  end for
  6.2 Add new selected member
  $$s_{iter} = \underset{k}{\text{argmin}}\ \overline{\text{corr}}_k$$
  6.3 Update variables
  $iter = iter + 1$, and $cm = \mathbf{Y} \backslash \mathbf{s}$
until $iter = nm$

It is important to note that, in the context of feature selection, the variables redundancy is measured from the absolute value of the correlation, as proposed by Haindl et al. [75]. However, in the selection context of a prediction ensemble, the procedure should encourage member selection with negative correlation. Furthermore, we propose a FGS instead of the BGS proposed by Haindl et al. [75].

Mutual information

In the previous method, supported by the analysis presented in Sect. 1.1.3, we established that the selected members must present negative or nil correlation in order to allow ensemble error reduction; however, this evaluation focused on the average linear correlation ignoring key aspects such as a possible nonlinear relationship (mutual information), the degree of correlation between the members themselves (redundancy), and the consideration of the observed data as an indicator of the relevance of each member.

Thus, we explore member selection using mutual information in the same descriptive framework proposed by Brown [37] (see Sect. 2.4.2), which is based on the maximization of the first order utility criterion:

$$J_k = \underbrace{I(\mathbf{y}_k; \mathbf{o})}_{\text{relevance}} - \beta \underbrace{\sum_{m=1}^{nv} I(\mathbf{y}_i; \mathbf{y}_m)}_{\text{redundancy}} + \gamma \underbrace{\sum_{m=1}^{nv} I(\mathbf{y}_k; \mathbf{y}_m | \mathbf{o})}_{\text{conditional redundancy}},$$

Several authors have proposed different criteria with various penalties to manage redundancy. Table 5.1 shows the various configurations tested in this study, more details on each criterion can be found in Brown [37] or in the references given in the last column. Note that the criterion CMIM evaluates the max operator between the two redundancy terms instead of the individual sum in Eq. 2.2.

Table 5.1: Parametrization proposed to evaluate the mutual information given by the Eq. 2.2.

Criterion	β	γ	Reference
Pure Mutual Information Maximization (MIM)	0	0	Battiti [18]
Mutual Information based Feature Selection (MIFS)	1	0	Battiti [18]
Maximum-Relevance Minimum-Redundancy (MRMR)	$1/(nv-1)$	0	Peng et al. [117]
Joint Mutual Information (JMI)	$1/(nv-1)$	$1/(nv-1)$	Yang and Moody [165]
First-Order Utility (FOU)	1	1	Kwak and Choi [94]
Conditional Mutual Information Maximization (CMIM)	1	1	Fleuret [61]

Evaluation of the terms of Eq. 2.2 requires the discretization of the information. For this, we consider quantile-based transformation taking into account nine scenarios ranging from 2 to 10 quantiles uniformly evaluated. So, if the information is discretized into 2 classes, the median will be the basis of the categorization of each variable.

Although the methods of selection based on mutual information are of the filter type, here we propose a linear search for the best combination of maximization criterion and the number of quantiles sets on the discretization, based on the minimization of the CC given by Eq. 5.2. Algorithm 4 presents the member selection scheme in which a FGS is applied in order to choose, at each iteration, the feature with the largest incremental gain, defined by one of the methods shown in Table 5.1.

Backward greedy selection

Hydrological members are sequentially removed from a candidate set of 800 members. The removal process runs until it achieves a minimum number of members or until the CC increases.

The configuration of this technique requires the definition of two subsets to run a cross-validation in order to avoid overfitting and the establishment of a minimization criterion. For selection of members in probabilistic prediction, it is necessary to define a minimum number of members to retain, reflecting the desired accuracy in estimating the predictive PDF.

Algorithm 4 Linear search of criterion and optimal discretization in MI selection.

1. Define the features space (members space):
   $\mathbf{Y} = \{y_1, y_2, \ldots, y_d\}$, where $d = 800$ in this case.
2. Define the number of members to select nm.
3. Define the methods to evaluate mutual information (see Table 5.1).
   here $methods = \{\text{MIM}, \text{MIFS}, \text{MRMR}, \text{JMI}, \text{FOU}, \text{CMIM}\}$
4. Define the number of quantiles to set on variables categorization.
   here $qt = \{2, \ldots, 10\}$

for all combinations of methods and number of quantiles **do**

5.1 Evaluate first selected member based on relevance

for all $\mathbf{y}_k \in \mathbf{Y}$ **do**

$I_k = I(\mathbf{y}_k; \mathbf{o})$

end for

$s_1 = \underset{k}{\text{argmax}}\ I_k$

5.2 Define initial candidate members set, $cm = \mathbf{Y} \backslash \mathbf{s}$, and initialize $iter = 2$

5.3 Evaluate the other $nm - 1$ members

repeat

for all $\mathbf{y}_k \in cm$ **do**

Evaluate the J criterion (Eq. 2.2)

end for

$s_{iter} = \underset{k}{\text{argmax}}\ J_k$

5.4 Update variables

$iter = iter + 1$, and $cm = \mathbf{Y} \backslash \mathbf{s}$

until $iter = nm$

end for

6. For each selection corresponding to each combination $sel = \{method_i, qt_j\}$ evaluate the CC (Eq. 5.2).
7. Evaluate best combination, $best\_comb$

$best\_comb = \underset{sel}{\text{argmin}}\ \text{CC}$

Expanding on the experiments performed in Chap. 3, we simplify the configuration of the training and validation subsets using a simple interleaving data: for every four consecutive data in the estimation dataset, the fifth is assigned to the validation dataset.

Nondominated sorting genetic algorithm II

This technique focuses directly on the search for the optimal Pareto front without a prior weight definition of the objective functions to be minimized. Like any evolutionary algorithm, its configuration is given by way of representing potential solutions (genotype), evaluation function (or fitness function), population, initialization, parent selection mechanism, variation operators, survivor selection, and termination condition. Table 5.2 presents a summary of the configuration evaluated.

Representation: each candidate solution is defined by the following guidelines:

Table 5.2: Description of the NSGA-II for the hydrological members selection problem.

Representation	Truncated permutations
Recombination	Partially mapped crossover
Recombination probability	90%
Mutation	Swap
Mutation probability	2% for each allele
Parent selection	Best 2 out of random 4
Survival selection	Pareto-front rank and crowding distance
Population size	100
Number of offspring	2
Initialization	Random
Termination condition	300 generations

- The individual should represent 48 members *mutually exclusive* in order to compare solutions with other methods.
- In terms of selection of variables, the permutation of the same group of members does not represent a new solution.

Thus, the representation of each individual (candidate solution) is a permutation of a set of 800 integers. However, only the first 48 alleles (string positions) represent the solutions to be tested. The other 752 alleles are reserved for the application of variation operators.

Fitness evaluation: As shown in Sect. 2.4.4, a crowding distance metric is defined for each point as the average side length of the cuboid defined by its nearest neighbours in the same front. The larger this value, the fewer solutions reside in the vicinity of the point [58]. The crowding distance in this case is defined by the two scores evaluated in this study: IGNS and RD_{MSE}.

Population and parent selection mechanisms: The population was set to 100 individuals, the mating population was set to 2, and there were two offspring. We use a tournament operator which considers first dominance rank, then crowding distance, more details can be found in Deb et al. [50]. Tournament selection involves randomly picking a number of strings from the population to form a "tournament" pool. The two strings of highest fitness are then selected as parents from this tournament pool.

Initialization: The initial generation starts by randomly initializing a population of points using Latin hypercube sampling [105].

Variation operators: It is clear that we need variation operators to preserve the permutation property that each possible allele value appears exactly once in the solution.

Following Vrugt et al. [151], the crossover rate was set to 0.9, while the mutation rate was set to $1/l_c$, where l_c is the length of chromosome, i.e. 48. In these experiments, the mutation

rate represents the probability that each gene in the chromosome will mutate and not the probability that a single mutation will occur in the chromosome.

Given the particularity of the representation in this problem, where the first 48 alleles define the solution to be evaluated and the 752 remaining ones serve as support for the variation operators, the following mechanisms are also exploited:

- Recombination: the Partially Mapped Crossover proposed by Whitley [159] is used. We consider the random choice of crossover points between alleles 1 and 47 for its implementation.
- Mutation: We use the swap mutation method [58], wherein the first position corresponds to one of the first 48 alleles and the second position corresponds to one of the last 752 alleles.

Survivor selection: The two populations are merged and fronts assigned. The new population is next obtained by accepting individuals from progressively inferior fronts until it is full. If not all of the individuals in the last front considered can be accepted, they are chosen on the basis of their crowding distance [58].

Termination condition: We define 300 generations or 30 000 fitness evaluations as the stop criterion.

Moreover, with the goal of comparing other techniques, a representative solution of the Pareto front is necessary, in which case we orient this selection with the post-Pareto front analysis proposed by Chaudhari et al. [42]. Therefore, the procedure contains the following steps:

- Obtain a sub-set of solutions that represent the Pareto-optimal front.
- Apply k-means clustering, so a normalization space is needed to avoid problems arising from the scale of the scores. Here we normalize each variable so that they will have zero mean and unity standard deviation. To find the "optimal" number of clusters, we evaluate the number of clusters that maximizes the mean silhouette value. The silhouette is a measure of how close each point forming a cluster is to others in the neighbouring clusters [101].
- For each cluster, select a representative solution. To do this, the solution that is closest to its respective cluster centroid is chosen as a good representative solution.
- Analyze the "knee" cluster or the k representative solutions. In this study the "knee" cluster represents the closest BGS model because we do not propose the scenario in which one could decide which of the two functions to minimize is more relevant at a given time.

5.3 Results and Analysis

The previous chapters showed that simplification of a HEPS may even lead to improved quality of the forecast; however, as discussed in Table 3.5, the simplification complexity is inversely related to the number of members to retain. In this context, Table 5.3 compares results of the 800-member reference HEPS and of the three schemes that allow to infer about

the complexity of the simplification process: 16 members (1 deterministic prediction and 16 hydrological models), 48UP (3 representative meteorological members and 16 hydrological models – uniform HMP), 48MR (median of 200 evaluations for the arbitrary choice of the number of meteorological members to propagate into hydrological models, respecting a scheme of 48 members). Note that these results differ slightly from those reported in the previous chapters due to differences in the predictive PDF evaluation (see Sect. 5.2.1).

Table 5.3: Probabilistic performance for the 9th FTH in original 800-member, 16-member and two 48-member HEPS schemes.

Training Catchments	HEPS members	Scores		Testing Catchments	HEPS members	Scores	
		RD_{MSE}	IGNS			RD_{MSE}	IGNS
H24	800	7.08	0.41	H36	800	3.50	− 0.09
	16	37.34	13.43		16	32.95	13.79
	48UP	8.20	− 0.49		48UP	5.00	− 0.75
	48MR	7.82	− 0.46		48MR	4.67	− 0.74
H93	800	2.59	− 0.27	M04	800	1.74	− 0.03
	16	27.99	12.01		16	25.44	12.15
	48UP	4.47	− 0.98		48UP	4.31	− 0.72
	48MR	3.84	− 0.95		48MR	3.50	− 0.70
M06	800	1.42	− 0.14	M15	800	1.61	0.28
	16	24.59	11.87		16	27.12	12.09
	48UP	4.25	− 0.77		48UP	3.37	− 0.64
	48MR	3.19	− 0.74		48MR	2.62	− 0.61
P70	800	4.14	3.29	P72	800	4.39	0.89
	16	31.65	19.30		16	31.97	18.03
	48UP	5.22	− 0.06		48UP	5.28	− 0.39
	48MR	4.51	0.06		48MR	4.51	− 0.33

First, the lesser 9th FTH performance of deterministic HEPS is quite obvious – a biased model (high IGNS) with inadequate variability (high RD_{MSE})[†]. In this regard, several authors have already highlighted the strong influence of the variability of the weather forecast on the HEPS diversity [83, 115, 145]. In this sense, Velázquez et al. [148] has demonstrated that it is possible to achieve further gain through the combination of probabilistic meteorological prediction and a group of hydrological models (conceptual diversity).

Second, with reference to the 48-member schemes, it is noteworthy that, in all cases, the IGNS of reference is improved, but, at the expense of reliability. Specifically, the uniform HMP scheme is the most efficient in minimizing IGNS, while the median of random evaluations is a little less inefficient in the reliability term (RD_{MSE}), but still do not match the 800-member HEPS performance. Consequently, simplification stands out as an optimization task involving hydrological models in a weight assignment problem. To get a clearer view of

[†]Note that a high RD_{MSE} may be associated with both high and low ensemble dispersion.

the members reduction in the simplification task, we show in Fig. 5.2 the normalized score evolution on both estimation and validation datasets when BGS technique is applied.

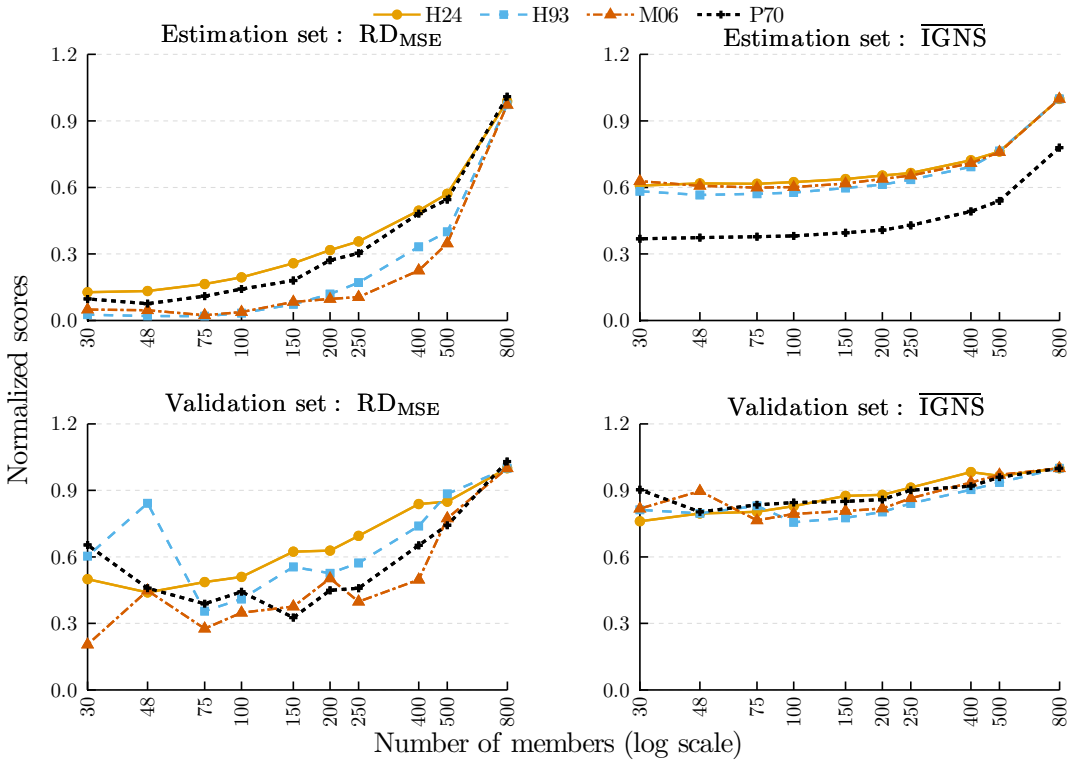


Figure 5.2: Behaviour of normalized scores in the BGS on training and validation datasets. Catchments H24, H93, M06, and P70.

At first it can be seen that the gain on the validation set is of lower magnitude than on the estimation set, which may be an indicator of structural differences between the two datasets. Ideally, these two sets must be "homogeneous" so validation results reflect the same phenomenon than for the estimation dataset, however, given the length of the series and method of datasets conformation, it is difficult to guarantee such a property. But, it can be seen that the validation results are always smaller than one, which indicates that the selection reflects also an optimization process. As expected, the estimation dataset shows smooth curves of the learning phase of the algorithm. Furthermore, validation conforms to what should be expected as generalization of the algorithm. In this case, it can be seen that the greatest gain is in minimizing RD_{MSE} because while normalized reliability score is between 0.45 and 0.82 for the 48-member HEPS, in this same situation the normalized IGNS shows greater values between 0.78 and 0.91.

Also, the simplification process on the validation set shows that the gain becomes unstable especially when the number of members is less than about 100. However, we can see that the selection of 48 members implies a gain or loss of quality not very significant with respect to the 100 members, except in basin H93 for the reliability case and basin M06 for the IGNS

case. So, a selection scheme of 48 members was arbitrarily established as a basis for techniques comparison. Additionally, 48 is a multiple of the number of available hydrological models (16), which facilitates the setup of the uniform HMP scheme.

Furthermore, note that while the application of BGS is based on minimization of the CC (Eq. 5.2), there is no strict internal control over the priority of each score despite the possible definition of the weights of each single score, because one can minimize one score causing the detriment of others.

In an effort to visualize explicitly the compromise between the scores, the BGS can be executed with different score weights in the CC definition, which leads to repeatedly evaluate the BGS at the expense of high computational cost. The computational complexity[‡] of the BGS technique is the order of $O\left(d^2\right)$, where d is the number of members of reference, in this case 800. Consequently, it may be more efficient to resort to a multiobjective technique such as NSGA-II, where optimization is centralized in the group of solutions that represent a compromise between the objectives evaluated in the so-called Pareto optimal front.

Figure 5.3 presents such a Pareto-type analysis with NSGA-II. Each panel illustrates the behaviour of different selections for each basin. Inset figures in the upper right corner of each panel show the 30000 tested selections in the optimization process. Symbols in bold represent the centroid of each cluster for catchments H24, H93, M06, and P70. The BGS solution is shown (cross) to allow a direct comparison. In these figures, we can see that the reliability of the system is the principal component of the variability of the results. Similarly, the density of points outside the bounding rectangle limited by unit normalized scores, is evidence that the optimization process converges rapidly in the initial performance of the reference HEPS. Thus, the simplification process is shown primarily as a process of scores optimization, hence it may be referred to as a post-processing process.

The Pareto front obtained in the optimization is drawn in each panel, along with their respective clusters and centroids identification. Additionally, the BGS solution is presented by a cross marker that exhibits scores slightly different from those shown in the estimation process in Fig. 5.2, because, at this stage, it was evaluated with a predefined number of members (48) and using all the basin information without cross-validation (subdivision of data). The scale of normalized scores of each panel is similar, except for basin P70, where the IGNS and the $\mathrm{RD_{MSE}}$ exhibit larger Pareto ranges, between 0.39 and 0.48 and between 0 and 0.6, respectively. This difference could be related to two factors specific to this basin: its smaller drainage area (see Table 3.2) and its higher relative streamflow.

In general, in Table 5.4 we can see that the BGS solution is acceptable. For example, in basins

[‡]To specify algorithm requirements, independent of the current computer hardware (which is always changing anyways), the big oh notation, $O(\cdot)$, is used to show an upper bound on the resources needed to solve a given problem [56].

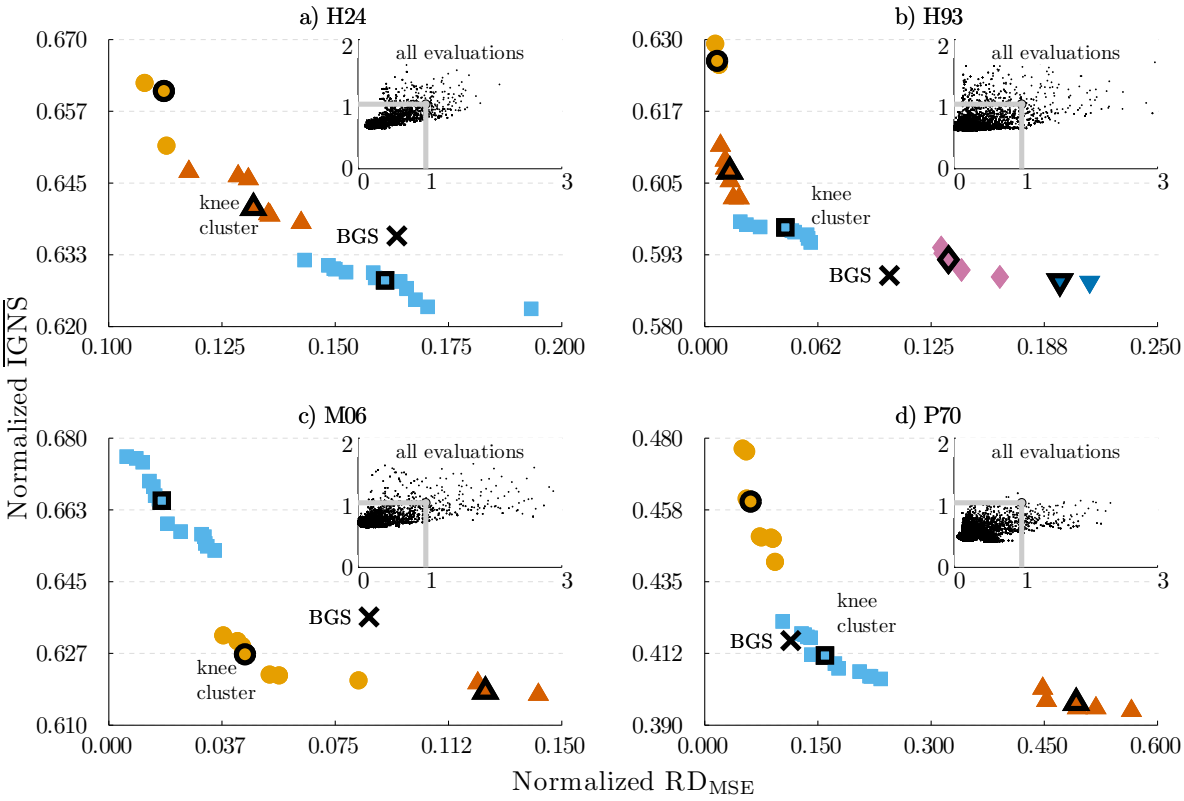


Figure 5.3: Evaluation of different selections with NSGA-II. Trade-offs between mean IGNS and RD_{MSE}.

M06 and P70, the BGS solution is part of the Pareto front, while in the other two basins, BGS selections belong to a front very close to the optimum found with NSGA-II, i.e. the "knee" cluster centroid (solutions in bold). Importantly, the Pareto front offers a descriptive version of the optimization process, which allows the development of a decision process based on the characteristics of each score and the properties we want to prioritize in a particular case. In other words, NSGA-II offers more flexibility to the operational hydrologists than BGS.

As explained in Sect. 2.1, a rigorous test requires testing the model, in this case of simplification, against new information. Thus, the above results are optimistic indicators. Accordingly, Fig. 5.4 shows the results of different selection techniques in a framework of extrapolation in both time and space – executed on a nearby catchment at different FTHs. Note that the y-axis represents the normalized scores, so in left panels y-axis corresponds to the normalized RD_{MSE} (reliability), while in right panels, it indicates the normalized IGNS (bias). Additionally, in the legend, UP represents the Uniform hydrological models Participation and bp-rand indicates the boxplots of the 200 random selections executed.

Regarding the reliability of the system, these results demonstrate again that the main difficulty lies in the preservation of this property. Furthermore, random selections show that the

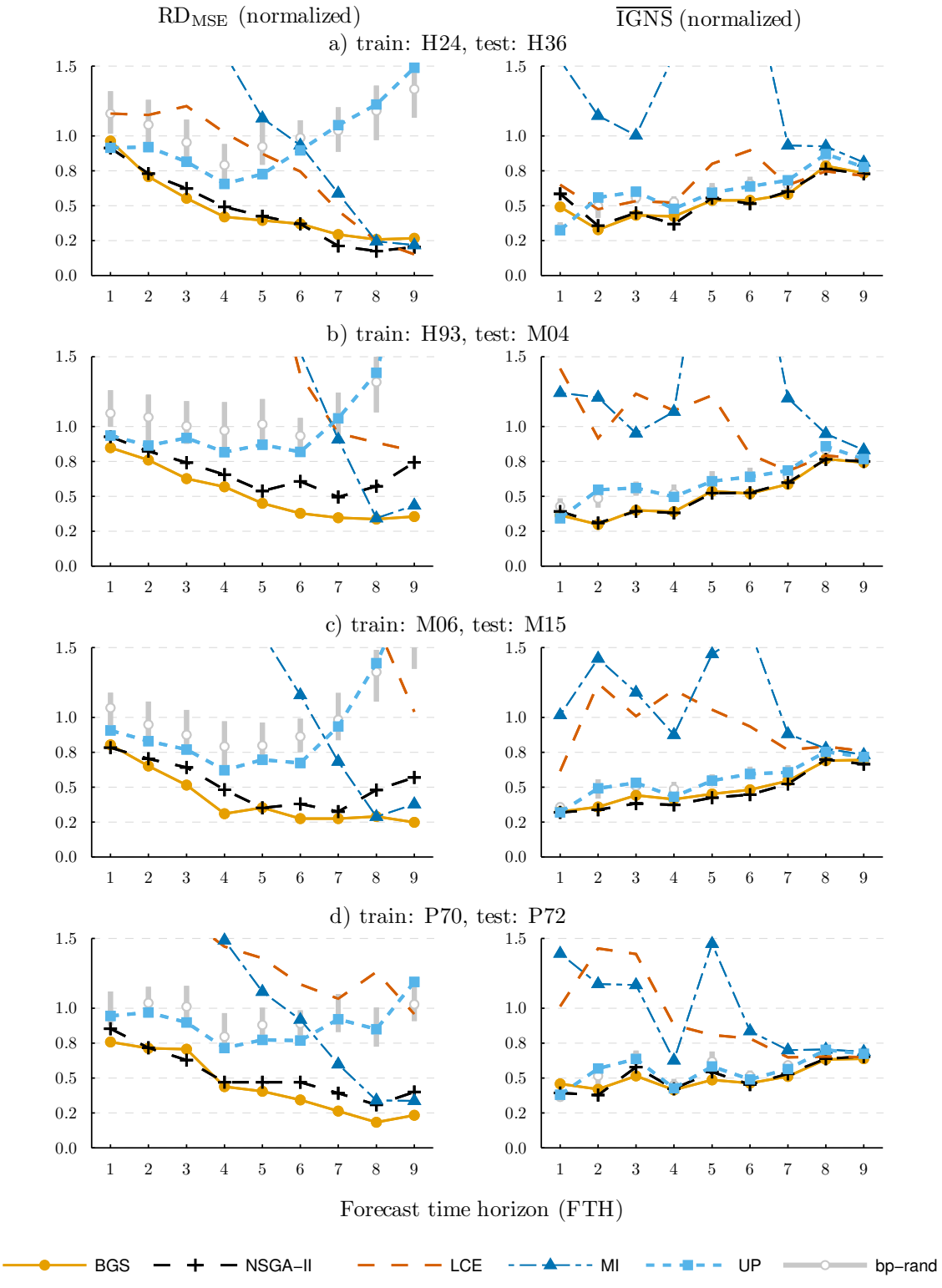


Figure 5.4: Comparison of different HEPS simplification schemes of 48 members.

Table 5.4: Normalized scores for the 9th FTH and 48-member HEPS schemes for the BGS and representative solutions of NSGA-II.

Basin	Selection techniques	Normalized scores		Basin	Selection techniques	Normalized scores	
		RD_{MSE}	IGNS			RD_{MSE}	IGNS
H24	BGS	0.164	0.636	M06	BGS	0.086	0.636
	NSGA II 1	0.161	0.628		NSGA II 1	0.018	0.665
	NSGA II **2**	0.132	0.641		NSGA II **2**	0.045	0.627
	NSGA II 3	0.112	0.661		NSGA II 3	0.125	0.618
H93	BGS	0.102	0.589	P70	BGS	0.114	0.417
	NSGA II 1	0.007	0.626		NSGA II **1**	0.159	0.412
	NSGA II 2	0.196	0.588		NSGA II 2	0.061	0.460
	NSGA II 3	0.014	0.607		NSGA II 3	0.492	0.397
	NSGA II 4	0.135	0.592				
	NSGA II **5**	0.045	0.597				

situation is more dramatic in FTH over 6 days, except in basin P72, where the median of the random selections is around one. This behaviour is especially important if one considers that the benefits of the 800-member HEPS is focused primarily on these FTHs. Also note that the interquartile range (length of the box plot) exhibits an important dispersion in relation to results observed with respect to IGNS. Concerning the uniform HMP scheme, it is important to note that the tendency is similar to the random selections; however, it is remarkable that, in the initial FTHs (1 to 6 days), uniform selection is generally better than the first quartile of the 200 random selections tested. In relation to the LCE, the normalized RD_{MSE} shows poor performance, except in the latest FTHs in basins P72 and M04. With regard to the selection with MI, this technique presents a good performance in the last three FTHs. Nevertheless, this technique rapidly loses the ability to generalize in the earlier FTHs. Note the similarity of the LCE and MI selections in basin H36, which indirectly leads to the inference that the correlation of this basin is approximately linear. In relation to BGS and NSGA-II, the proximity and high performance achieved in both techniques is obvious.

Regarding IGNS, the relationship between the selections guided by correlation and the ensemble bias is confused. In general, the LCE technique is more efficient than the MI technique, which can be attributed to the lack of an explicit formulation aimed at selecting members with negative correlation in the methods used in the MI selection. However, except for basin H36, LCE performance is much lower than those reported by BGS and NSGA-II, which show very high efficiency (low normalized scores). At this point, it is important to note the inverse relationship between gain and FTH – shorter FTHs are generally followed by a higher gain. Furthermore, note that the box plots for random selections and uniform HMP show low dispersion and a lower performance than BGS and NSGA-II.

Finally, Fig. 5.5 compares BGS and NSGA-II HMP. The similarity of these solutions is obvious, which confirms the importance of models HM#3 and HM#14 in these four basins. Such a combination of models is interesting given that HM#3 and HM#14 present opposite performance values, as discussed in Sect. 4.1.2, in terms of diversity evaluation (mean performance rank), where HM#3 stands out as one of the best performing models, while HM#14 shows poor performance.

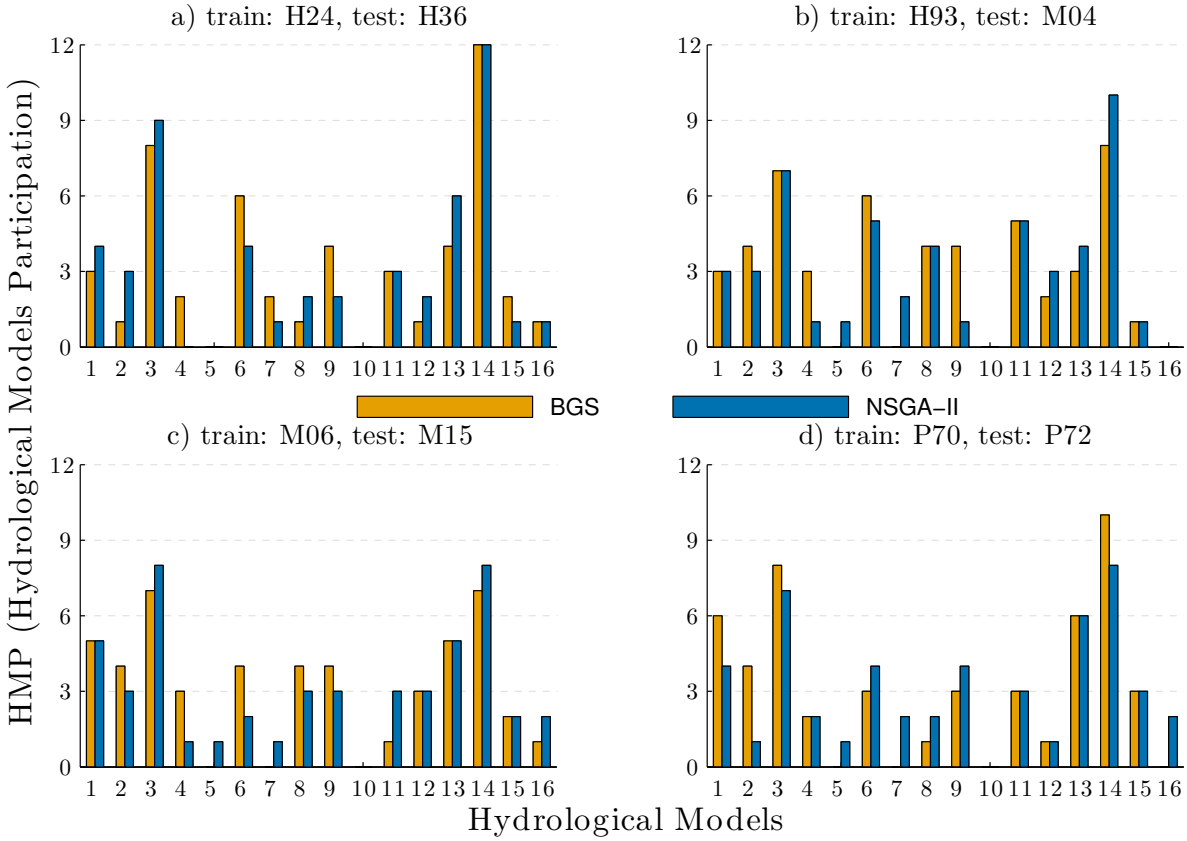


Figure 5.5: Comparison between HMP results of BGS and NSGA-II.

5.4 Conclusion

Given the singularity of HEPS in evaluation, where its 800 hydrological members come from the propagation of 50 interchangeable meteorological members, simplification scheme and/or scores optimization of the system based on the HMP has proven to be highly efficient. Clearly, the methodology presented here combines HMP and meteorological clustering stage as additional filters that facilitate the interpretation of the hydrological member selection. However, in the case of a HEPS conceived from non-interchangeable meteorological members (in Canada, for example), the selection task would then directly identify the importance of certain members in the propagation of uncertainty in streamflow prediction.

An intuitive choice for simplifying the 800-member HEPS is supported by the correlation of hydrological members, i.e. members who provide the same type of forecast are excluded as proposed in the LCE technique. Such scheme is consistent with the theoretical developments in terms of reducing the bias of the system, which can be corroborated in the neighbour extrapolation for the eighth and ninth FTH. However, it fails at generalization in the other FTHs, which is an undesirable effect in a single simplified scheme.

In fact, it is important to highlight the structural deficiencies of the two correlation schemes proposed. On the one hand, simplification based on the LCE assumes the mean correlation as a "piecewise" selection criterion ignoring redundancy with respect to each other. In contrast, the MI selection explicitly proposes the selection of members individually relevant and not redundant with respect to each other.

With respect to the reliability of the system, this property is notoriously the most difficult to maintain in the simplification scheme. The results of a uniform HMP and random selections show such a difficulty. In this work, without loss of reference bias, it is important to note the high efficiency of BGS and NSGA-II, which we consider to be the best choice among the techniques evaluated. At this point, operational hydrologists can choose either taken into account such aspects:

- *Computational complexity*: Although BGS is more intuitive, BGS complexity is the order of $O\left(d^2\right) = O\left(800^2\right)$, while NSGA-II has a computational complexity of $O\left(s \cdot p^2\right) = O\left(2 \cdot 100^2\right)$, where s is the number of scores and p is the population size. However, in our experiments, the average running time shows that NSGA-II is about 5 times faster than BGS[§]. It is worth noting the availability of free software for the implementation of NSGA-II [62].
- *Trade-off between bias and reliability*: Although it is possible to run BGS multiple times with different score weights to display the trade-off between bias and system reliability, this procedure does not guarantee convergence to the Pareto optimal set [99] and increases the computational complexity cited above. In this sense, NSGA-II shows directly the different simplification schemes that highlight the trade-off among the evaluated scores.
- *Optimization of the number of members to retain*: Because the objective of our methodology is focused on techniques comparison with the same number of members, solution representation (genotype) with NSGA-II was established by permutations However, a binary encoding is more intuitive in order to optimize at the same time the number of members to hold out. In BGS, such an analysis is straightforward on the selection performance curve.
- *Search procedure*: BGS is a local search procedure and does not guarantee finding the optimal subset. In opposition, the NSGA-II has theoretically the capability to find the

[§]We use MATLAB as language, GNU/Linux Ubuntu operating system and a computer with Intel Core i7 920 2.7GHz CPU and 12GB of RAM. With these specifications, it takes about five hours to complete one run of BGS, while one NSGA-II run takes about 1 hour.

global optimum in its evolutionary search process, although there is no guarantee that it will find the global optimum.

Finally, we propose various directions for future research works:

- Further research with longer databases is needed in order to identify the HEPS value in several types of events, e.g peak events.
- Evaluate in a larger number of basins the relationship between the performance of the hydrological members selection and physiographic and hydro-climatological properties.
- Furthermore, diversity evaluated from the deterministic performance of each model, should be considered as an approximation of the true structural diversity of hydrological models. In this sense, an explicit analysis of the relationship between the structural diversity of a group of hydrological models and their relevance in a probabilistic scheme should be studied in more detail.

Part III

ANN Ensembles as HEPS

Chapter 6

Diversity from Dataset and Parametric Levels

In this chapter, we evaluate an ensemble model formed by 30 Feed-Forward Neural Networks (FFNNs) for predicting daily streamflows. We focus on diversity imposed by training each FFNN (ensemble member) with different subsets of information. With this objective, we propose the use of a clustering technique to select representative input vectors for training, which is known as stratification or stratified sampling.

The time series used in this study correspond to 12 basins evaluated in the second and third workshop of the MOdel Parameter Estimation eXperiment (MOPEX) project, which are freely distributed. These basins represent different hydroclimatological regimes.

Although the ultimate goal of our work is to establish the importance of diversity in the formation of a Hydrological Ensemble Prediction System (HEPS), in this chapter, we present only deterministic results to assess the impact of stratification in the FFNN ensemble training.

It is important to highlight that we present stratification skill using as a baseline an Ensemble of 30 FFNNs trained with early stopping using a **R**andom sampling of **100 P**ercent of the available information and a single predefined set of inputs variables (R100P). Thus, although neither the baseline model or stratified schemes represent the confluence of the latest advances in some Artificial Neural Network (ANN) topics as data selection, Input Variable Selection (IVS), training algorithms, and ANN structures, results of both models allows us to establish the importance of the "ensemble methods" with respect to other ANN architectures.

6.1 Stratification concept for ANN training

ANNs are characterized by their high interpolation ability, in contrast to their poor extrapolation capacity, i.e. if the training data does not contain the maximum possible output values,

an unmodified ANN will be unable to simulate the peak values [80]. Another crucial aspect common to most models that calibrate its internal parameters with respect to a desired or observed value (supervised models) is the lack of generalization, i.e. the deficiency to reproduce events not evaluated in the training phase. However, this flaw is often restrained using a technique called cross-validation, which exploits three subsets: one for estimating the parameters of the ANN fitting (estimation subset), a second one to check the performance (validation subset) in the training phase, and a final one to generate the expected generalization error (test subset).

In this order, interpolation capability and overfitting are strongly related to a correct definition of the cross-validation subsets. This scenario highlights the concept of stratified sampling or representative partitioning. The latter refers to the evaluation of the patterns contained in the information for guiding a subsequent resampling including different types of data.

Consequently, in the literature, different solutions have been proposed based on clustering tools. But new questions emerge about the clustering model and its corresponding configuration. Additionally, two other recurrent methodological issues persist: the definition of the space to clustering and the distribution technique from data clusters to subsets.

These questions have been explored by several authors from different types of view. For example, Anctil and Lauzon [8] proposed the clustering of the input space of the ANN with a Self-Organizing Map (SOM) technique to obtain various kinds of events, and then the construction of a sample of vectors for the cross-validation subsets is accomplished by sampling equally from each of the classes of vectors. Shahin et al. [134] also evaluated the SOM model but taking into account both the input and the output space. In this case, the data distribution was oriented in terms of the Kohonen layer configuration and availability and priority of information for estimation, testing, and validation datasets. In this same study, a stratification was proposed based on the fuzzy clustering algorithm, for which the data within each cluster are ranked in accordance with their degree of membership. Next, each data point is assigned to one of ten equally spaced membership intervals, one data point from each membership interval is assigned to the validation dataset, and another data point from that interval is assigned to the testing dataset while the remaining data points from the same interval are assigned to the estimation dataset.

Recently, May et al. [103] evaluated different manners in which samples are selected from SOM units: equal allocation, proportional allocation, and Neyman allocation, highlighting the reliability of the latter compared to the others. In this same way, Bowden et al. [25] showed the efficiency of the identification of patterns combining SOM with nonparametric kernel density estimators to calculate local density estimates, with the aim of identifying the model's range of applicability and assessing the usefulness of the forecast.

In summary, in hydroinformatics applications, the use of SOM in the stratification is more

popular than other clustering methods; given that it is another class of ANN, it also may already be relatively familiar to ANN modellers [103]. Here, we adopt a methodology of stratification based on k-means clustering given its simplicity, speed, and relative stability. Such a methodology was presented initially by Diamantidis et al. [51] for classification problems and later adapted by López et al. [96] for forecasting problems. In this methodology, the clusters are evaluated in the input or the output space. Regarding the distribution of the data from each cluster into cross-validation subsets, the distance between each event and the centre of its respective cluster is the basis of the data allocation, ensuring that cross-validation subsets retains the spatial distribution of the data in each cluster.

Additionally, we evaluate the influence of the number of samples and type of stratification in the deterministic performance of the ensemble, guided by investigations from several authors [2, 8, 51, 96]. Regarding the number of examples, Domingos [52] showed that, while the ratio of algorithm accuracy with the expected error can be of logarithmic order, the ratio of the number of hypotheses to be tested is doubly exponential with regards to the number of inputs, a major problem known as "the curse of dimensionality". Also, another approach suggests an asymptotic limit error when the number of examples tends to infinity; however, the bias variance dilemma discussed in Sect. 1.1.2 shows that if model A is better than model B given an infinite number of data, B is often better than A given a finite dataset.

Several authors have specifically evaluated the relationship between the number of examples and performance in the domain of ANN [11, 96], concluding that, in general, a greater amount of information leads to better results in generalization – these experiments are usually accompanied by training methods such as early stopping or Bayesian regularization in order to avoid overfitting. At this point, we evaluate different scenarios searching a stratification methodology that allows us to extract sub-samples to train the network without sacrificing performance. It is important to note that the choice of calibration data in hydrology is recognized as a crucial aspect. Research into data requirements had led to the understanding that the informativeness of the data is far more important than the amount used for model calibration [72, 91, 140, 168].

6.2 Methodology

As mentioned above, the goal of this chapter is to evaluate the effect of different stratified training datasets on the deterministic performance of a stack of 30 FFNNs. So, Sect. 6.2.1 details the concept of stratification based on k-means to define the estimation and validation datasets. Note that we use a split sample approach to partition the samples into training and test datasets for a rigorous evaluation of the ANN model, ensuring the independence of the data used in the testing phase. The basic configuration of each FFNN is described in Sect. 6.2.2, following the guidelines given by various authors [1, 97] to ensure that the results

are repeatable and reproducible – coinciding with the postulates of the scientific method. Finally, in Sect. 6.2.3, deterministic performance functions used in this and the next chapter are presented.

6.2.1 k-means–based stratified sampling

Stratification is a resampling technique used to prevent unbalanced selection of calibration datasets. In the case of ANN, this task is often performed randomly or arbitrarily, which, in some cases, may lead to a poor performance. Here, we adopt the stratification method proposed by López et al. [96] with slight modifications. This method involves the division of data into smaller mutually exclusive groups. Each group or fold contains representative sub-samples. Thus, if we require a stratified sub-sample of 50% of information, the data should be partitioned into two folds, and if we require 25%, the data should be partitioned into four folds, i.e. the number of folds is approximately equal to $\lfloor 1/\text{percentage to sampling} \rfloor$*. Then, the modeller is free to choose one of these folds.

Once the inputs and outputs of the model are defined, it is necessary to determine the space to stratify. In pattern recognition, the output space is generally accepted as the domain of stratification [51]. We propose three schemes of stratification: one which includes only the input space (I), one that includes only the output space (O), and finally another integrating both the input and output spaces (I+O). Also, we evaluate four different percentages of resampling, corresponding to 12.5%, 25%, 50%, and 100% of the available training information.

The mechanics of stratification depends on three main modules: a clustering module which defines the data belonging to each cluster, a check module which assesses the relevance of the clustering based on the number of folds to form, and an allocation module which distributes data from clusters to folds.

With respect to the clustering module, we select the k-means method based on the euclidean distance (see Sect. 2.2.1). The correct choice of the number of clusters is often defined by trial and error, using measures such as the maximization of the mean value of the silhouette function [101]. In our case, the number of clusters (k) is initialized arbitrarily to 6, since in most cases the silhouette function analysis led to only two clusters, which is not more informative given the high dispersion in those two clusters.

Regarding the checking module, the condition imposed for the posterior allocation is that the number of data in each cluster (n_k), which must be higher or equal than the number of folds to set in the stratification ($n_k > \lfloor 1/\text{percentage to sampling} \rfloor$). This "sufficiency condition" allows that each one of the mutually exclusive sub-samples will contain at least one example of each cluster.

*$\lfloor a \rfloor$ is the "floor" function. It returns the largest integer smaller than a.

Finally, the allocation module or distribution of data from clusters to folds is responsible for the homogeneous distribution of the data points. Homogeneous distribution is achieved based on the distance of each data point to its respective cluster centroid, as proposed by Diamantidis et al. [51]. In this case, the data points are sorted based on such distance and subsequently distributed into the folds sequentially. For example, Fig. 6.1 shows that after clustering, data points are ordered as follows: {2,1,4,5,6,3}. Therefore fold 1 will contain the data point closest to the cluster centre, i.e. data point 2, the second closest data to the centre is assigned to the second fold, and so on, repeating the task allocation in a circular manner. So, after the allocation of data point 5 to the fourth fold, data point 6 is picked up and assigned to the first fold, while data point 3 is assigned to the second fold.

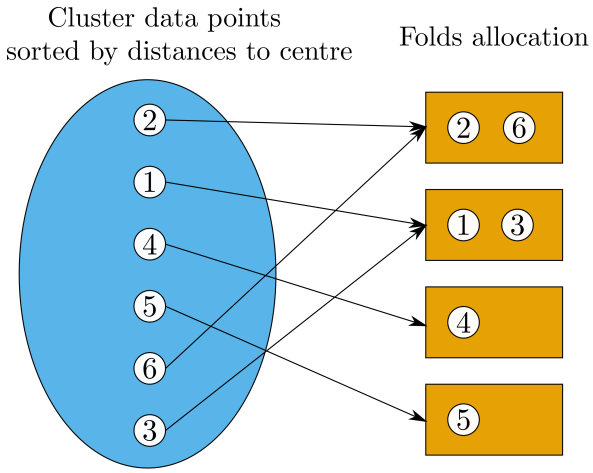


Figure 6.1: Assignment of similar data points to different folds.

The proposed methodology is based on the following steps (Fig. 6.2):

1. Define the number of clusters or determine it using an objective function, e.g. maximizing the mean silhouette value [101].
2. Determine the clusters and their respective centres.
3. If stratification seeks configuring two datasets with different percentages of resampling, as it is often the case defining the estimation and validation datasets within the ANN framework, the process begins with the most restrictive dataset, i.e. the one with the lowest percentage of resampling forcing the formation of a greater number of folds and therefore more data points by clusters. We evaluate different percentages for training ($p_{training}$), but in all cases the estimation ($p_{estimation}$) and validation ($p_{validation}$) percentages are defined as 75% and 25% of the training dataset respectively.
4. If the configuration of a validation dataset is needed, the sufficiency condition is verified ($n_k >= \lfloor 1/p_{training} \times p_{validation} \rfloor$). Otherwise, the process continues in step 7 intended to set the estimation dataset. If the sufficiency condition is not reached, the number of clusters decreases systematically and returns to step 2.

5. If the sufficiency condition is reached, the process continues with the distribution of data from clusters to folds, as shown in Fig. 6.1.
6. Clusters information is updated by removing values used in the configuration of the validation subset.
7. The sufficiency condition is verified ($n_k >= \lfloor 1/p_{training} \times p_{estimation} \rfloor$).
8. If the sufficiency condition is reached, the process ends with the distribution of data from clusters to folds to configure the estimation subset, otherwise the number of clusters decreases systematically and returns to step 2 to start the process again.

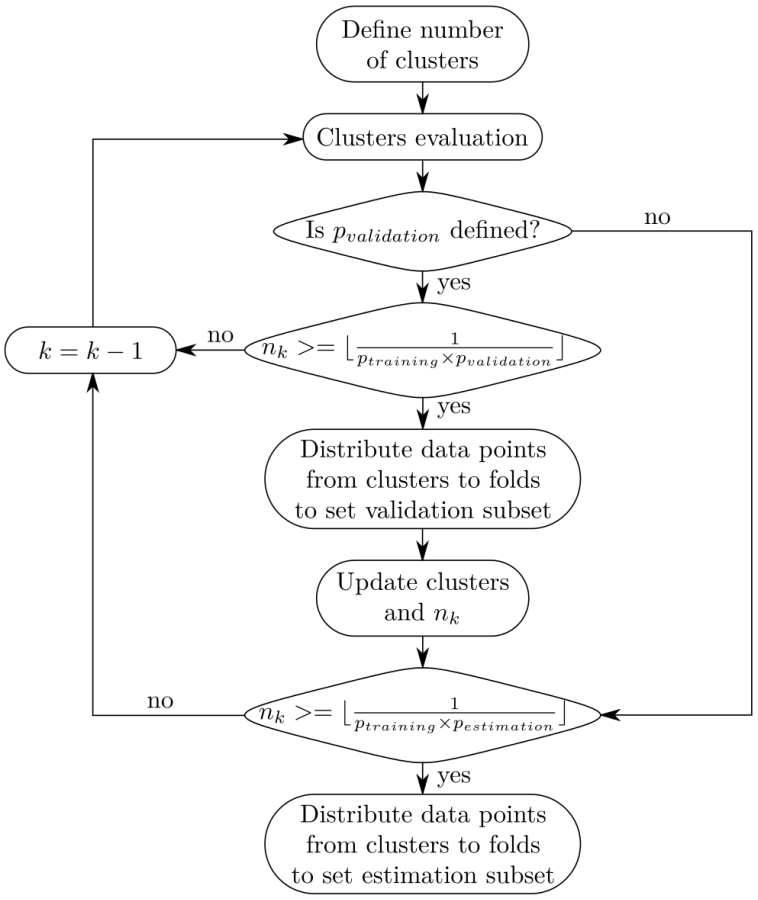


Figure 6.2: Flowchart of the proposed stratification methodology.

Now, consider the example shown in Table 6.1 and Fig. 6.3. We present a hypothetical model for the prediction of the streamflow (Q_t) based on the previous values of precipitation (P_{t-1}) and streamflow (Q_{t-1}). The goal is to obtain a stratified sample of 25% of the information based on the I+O space (4 folds). In this case, the restriction of the number of data per cluster (minimum 4) leads to gradually reducing the number of cluster from 6 to 2.

Note that to evaluate properly the k-means, one has to make sure that all dimensions have the same scale, so we normalize inputs and targets to have zero mean and unity variance. However,

Table 6.1: Stratification example. Bold events or data-points correspond to clusters represented by squared markers in Fig. 6.3.

id	Real scale			Standardized scale			id	Real scale			Standardized scale		
	Q_t	Q_{t-1}	P_{t-1}	Q_t	Q_{t-1}	P_{t-1}		Q_t	Q_{t-1}	P_{t-1}	Q_t	Q_{t-1}	P_{t-1}
1	**32.2**	**11.4**	**116.1**	**3.5**	**1.9**	**2.6**	15	5.3	1.2	28.1	−0.4	−0.5	0.0
2	**20.4**	**14.0**	**61.1**	**1.8**	**2.5**	**1.0**	16	6.4	4.7	23.9	−0.3	0.4	−0.2
3	**28.7**	**7.6**	**64.1**	**3.0**	**1.0**	**1.1**	17	0.1	0.1	0.0	−1.2	−0.7	−0.9
4	**20.7**	**0.2**	**93.8**	**1.8**	**−0.7**	**2.0**	18	0.4	0.7	4.4	−1.1	−0.6	−0.8
5	**19.1**	**1.1**	**84.5**	**1.6**	**−0.5**	**1.7**	19	1.5	1.6	4.1	−1.0	−0.4	−0.8
6	**24.8**	**1.5**	**73.7**	**2.4**	**−0.4**	**1.3**	20	0.3	0.3	1.4	−1.1	−0.7	−0.9
7	6.1	5.2	22.6	−0.3	0.5	−0.2	21	0.1	0.1	0.3	−1.2	−0.7	−0.9
8	5.7	0.5	43.1	−0.4	−0.6	0.4	22	0.5	0.6	0.4	−1.1	−0.6	−0.9
9	5.5	15.0	9.9	−0.4	2.8	−0.6	23	1.2	1.4	0.9	−1.0	−0.4	−0.9
10	8.4	4.2	29.4	0.0	0.2	0.0	24	0.2	0.2	2.0	−1.2	−0.7	−0.8
11	3.2	1.3	26.1	−0.7	−0.4	−0.1	25	0.4	0.4	4.7	−1.1	−0.6	−0.8
12	5.4	2.1	23.5	−0.4	−0.2	−0.2	26	1.1	1.1	4.6	−1.0	−0.5	−0.8
13	9.2	4.4	25.3	0.1	0.3	−0.1	μ	8.2	3.1	29.5	0.0	0.0	0.0
14	5.7	0.6	18.8	−0.4	−0.6	−0.3	σ	9.7	4.3	32.9	1.4	1.0	1.0

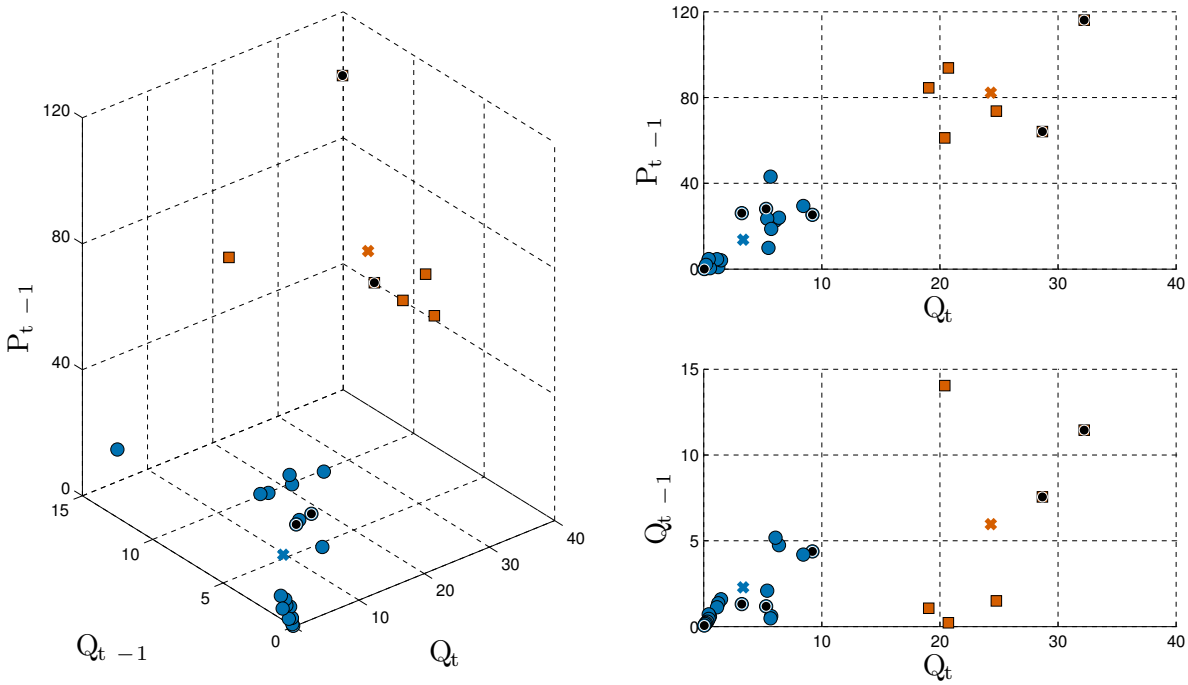


Figure 6.3: Space to stratify. Crosses represent the cluster centroids, circles and squares define two clusters distribution, and markers with a dot inside identify a selected stratified sub-sample.

when we evaluate the I+O scheme, to estimate the euclidean distance, it is imperative to provide the same level of importance to the single variable of the output space and the "d" variables of the input space. Consequently, the single normalized variable of the output space is multiplied by $\sqrt{d}$, in this case, $\sqrt{2}$. Note that the standard deviation of normalized output equals $\sqrt{2}$, i.e. the output space multiplier factor.

Data selection within each cluster is based on the distance between each instance and its cluster centroid. Thus, the instances are sorted according to these distances and the data are chosen for each fold with a mechanism showed in Fig. 6.1. Table 6.2 shows the process of selecting data for the first sample (first fold). From cluster 1, examples 2 and 6 are selected, and from cluster 2, examples 9, 7, 24, 18, and 26 are selected. Note that, if the number of data in the cluster is not much larger than the number of folds, we can get an unbalanced sample that somehow modifies the participation of the cluster in the problem. For example, cluster 1 contains 23% (6/26) of the data, but in the final sample, the participation of this cluster is 28% (2/7). Figure 6.3 shows the final selection of data in each cluster.

Table 6.2: Example of fold data distribution in cluster 1.

Example id.	Q_t	Q_{t-1}	P_{t-1}	Distance to centroid	Folds distribution
2	1.79	2.54	0.96	2.07	1
1	3.51	1.94	2.63	2.01	2
4	1.83	−0.68	1.95	1.48	3
5	1.59	−0.48	1.67	1.38	4
6	2.42	−0.38	1.34	1.08	1
3	2.99	1.03	1.05	0.92	2
Centroid	2.36	0.66	1.60		

6.2.2 ANN stack setup

Streamflow prediction models are configured with 30 FFNNs operating in parallel, yielding 30 responses at each time-step. Each FFNN is trained with stratified data from the methodology presented in the previous section. In this case, the lower the percentage of resampling data, the lower the probability that the FFNNs share the same training information. Additionally, the proposed stratification methodology encourages variety in the stratified datasets since k-means is evaluated with a random centre initialization and since the choice of stratified fold, between the $\lfloor 1/p_{training} \rfloor$ folds configured, is random.

We evaluate 30 experiments for each of the sixteen scenarios resulting from the combination between four percentages of data used to train an ANN stack (12.5%, 25%, 50%, 100%) and four resampling techniques (random and the three stratification schemes that differ by their application space). Additionally, the combination function, which represents the stack from

the deterministic point of view, is the average. Below we present the basic configuration of the FFNN used in the prediction ensembles.

Subsets definition

The series used in this study extending over 40 years (from 1960 to 1990), with a 50%-50% split sampling to determine the training and test datasets. Years 1980-1990 represent the available training information, while the first 19 years (1960-1979) are used for testing. This option at least guarantees the temporal independence of datasets information. It is important to highlight that a rigorous evaluation of the model should ensure the independence of the data used in the test [7].

Pre and post-processing

To warrant that all inputs and the output are on the same scale, they are linearly standardized so that their mean is zero and their standard deviation is one. It is thus clear that a reverse mapping of the network output is needed for comparison with the observed streamflow.

Network architecture

There is a wide range of applications of ANN architectures in streamflow forecasting. In this regard, Maier et al. present a detailed review of different architectures and experimental protocols [97, 98], which highlights the popularity of feed-forward ANN despite the potential benefits of using recurrent networks [49]. Additionally, recent advancements in ANN modelling revealed the high performance of Echo State Networks (ESNs) [48, 149]. Indeed, Vos [149] presents a comparison of various architectures of ANN with the same database evaluated here. Nonetheless, we move away from this trend of incremental technical refinement and promote the use of a simple FFNN with a single hidden layer trained using the Back-Propagation Levenberg Marquardt algorithm. ANN geometry was optimized by trial and error in the training dataset, resulting in six hidden neurons, which coincides with the evaluation of Anctil et al. [11], showing that there is not significant gain in using a higher number of hidden neurons. We use a hyperbolic tangent sigmoid as activation function in the hidden layer, whereas output neurons use a linear function.

Input space definition

As the object of this chapter is to assess the impact of the variety in the training sets in the calibration of an ensemble of ANNs, this aspect will be evaluated in more detail in the next chapter.

However, we use the proposed stratification to extract 50% of the data and determine the most relevant input variables with the FGS algorithm.

Various test showed that, in general, the preceding streamflow (Q_{t-1}) and precipitation (P_{t-1}) are the most important variables; nevertheless, depending on the analyzed series, other lags for these variables are highlighted in a second level of importance. Consequently, we adopted an integration scheme for all the series analyzed in which the following variables were grouped as a predefined input space: $Q_{t-1}, Q_{t-2}, P_{t-1}, P_{t-2}, P_{t-3}$, where "t" represents the time step for prediction purpose.

Optimisation set-up

The FFNN training is performed in batch mode, i.e. the parameters (weights and biases) are updated only after the evaluation of the whole training data (epoch). The initialization of the weights and biases follows the Nguyen-Widrow procedure [113]. This initialization method draws values in order to distribute approximately evenly the active region of each neuron in the layer across the layer's input space. The values contain a degree of randomness, so they are not the same each time this procedure is called. As mentioned above, we use the Levenberg-Marquardt back-propagation algorithm. We also use an adaptive learning rate initialized to 0.005. This value is multiplied by 0.1 whenever the performance function is reduced by a step. It is multiplied by 10 whenever a step would increase the performance function. With regard to the maximum number of epochs, this is set to 50; however, it is the early stopping method that ultimately governs the final epoch, which is around 20. Table 6.3 presents a summary of the ANN used in all experiments evaluated in this chapter. Our implementation is based on the Neural Networks Toolbox 7 of Matlab.

6.2.3 Performance evaluation

We use four deterministic measures: MAE, Mean Square Error (MSE), and two of its normalizations usually employed in hydrology: Nash-Sutcliffe Efficiency (NSE) criterion and Persistence Index (PI):

$$\text{NSE} = 1 - \frac{\text{MSE}}{\frac{1}{N}\sum_{t=1}^{N}(\bar{y}_t - \bar{o})^2}, \tag{6.1}$$

$$\text{PI} = 1 - \frac{\text{MSE}}{\frac{1}{N}\sum_{t=2}^{N}(\bar{y}_t - o_{t-1})^2}, \tag{6.2}$$

where N represents all the observed data points, $\bar{y}_t$ indicates the mean prediction ensemble at time t, o_t and $\bar{o}$ indicate the observation at time t and the mean observed value, respectively. These two dimensionless measures provide an overview of the model performance independent of the units or characteristics of the problem (e.g. the size of the basin or streamflow regimes). The PI (Eq. 6.2) offers a valuable alternative to (Eq. 6.1) by using the last observation (o_{i-1}) as prediction for all time steps. The PI statistics are particularly well designed for prediction evaluation, considering that the last observed streamflow is generally one of the ANN inputs. A negative PI value indicates that the model is degrading the provided information [11]. Both

Table 6.3: Neural network set-up.

Data preparation	
Preprocessing	Normalize inputs and targets to have zero mean and unity variance
Inputs	Predefined: $Q_{t-1}, Q_{t-2}, P_{t-1}, P_{t-2}, P_{t-3}$
Output	Q_t
Data division	Simple cutoff, training (years 1980–1990) and testing (years 1960–1979)
Network configuration	
Connection type	Feedforward
Geometry method	Trial and error
Geometry	5 inputs, 6 neurons in the hidden layer and one output (5-6-1)
Transfer functions	Hyperbolic tangent sigmoid (hidden layer) and linear function (output layer)
Training	
Style	Batch training
Initialization weights	Nguyen-Widrow initialization algorithm [113]
Training algorithm	Levenberg-Marquardt backpropagation algorithm [74] with early stopping
Learning rate	Adaptive, start with 0.005 and it is divided/multiplied by 10 if the error function decreases/increases
Max. Number of epochs	50
Stopping criterion	
Cross-validation	Validation performance has increased more than six times since the last time it decreased
Epochs	The maximum number of epochs is reached
Performance	Performance is minimized to the goal or the performance gradient falls below 1×10^{-10}
Learning rate	Learning rate becomes larger than 1×10^{10}

measures range from $-\infty$ to 1. These measures reach 1 for a perfect fit between predicted and observed values and 0 when the hydrological model is no better than a one-parameter 'no-knowledge' model [10, 11].

Skill based on a naïve model

We define as a baseline (naïve model) an R100P. In this model, as for all FFNN ensembles evaluated here, the FFNN configuration of each of 30 members ensemble corresponds to the structure described in Table 6.3. Then, we propose a skill or gain measure based on the median error of 30 evaluations of the stratified schemes (me_{scheme}) and the median error of 30 evaluations of the R100P model (me_{R100P}):

$$G_{scn} = \begin{cases} \dfrac{me_{scheme} - me_{R100P}}{me_{R100P}} & \text{when NSE or PI is used} \\ \dfrac{me_{R100P} - me_{scheme}}{me_{R100P}} & \text{otherwise} \end{cases} \tag{6.3}$$

A positive index indicates superior performance of the stratified schemes. The median error is used as a measure of central tendency due to the asymmetry of errors, especially in the random cases.

Note that the only difference between the R100P model and the other evaluated schemes is the selection of the training data, so results show directly the impact of stratification in the design of the ensembles. In this sense, the stratification is also evaluated in terms of clustering space and the percentage of stratified data.

It is important to highlight that the baseline model used as reference does not represent the confluence of the latest advances in some ANN topics as data selection, IVS, training algorithms and ANN structures, but we will show that the simple use of the "ensemble approach" with FFNN is as efficient as any other sophisticated ANN architecture.

6.3 Study area

We evaluate the daily forecast for the twelve basins used in the second and third MOPEX workshops [53], in order to exploit the availability and quality of the information from this experimental database[†]. Table 6.4 shows the nomenclature and a brief description of the properties of each basin. Having an average length of the series of 38 years and a spatial distribution of the basins in the southeastern U.S. allows the evaluation of different hydrometeorological conditions (Fig. 6.4), as indicated by the annual precipitation (P), streamflow (Q), and the precipitation to potential evapotranspiration ratio ($\overline{\mathrm{P}}/\overline{\mathrm{PE}}$) – reciprocal ratio to aridity index. A high $\overline{\mathrm{P}}/\overline{\mathrm{PE}}$ ratio indicates wet climate and a low value, dry climate. Mean values represent the mean annual values from 1960 to 1998. Actual evapotranspiration is represented by the difference between the mean precipitation and the mean streamflow ($\overline{\mathrm{AE}} = \overline{\mathrm{P}} - \overline{\mathrm{Q}}$).

Figure 6.4 shows that the experimental relations follow the trend of the theoretical formulations of Schrieber, Ol'dekop, and Turc-Pike about the hydrological characterization of the basins, such formulations can be found on Arora [16]. It is important to retain that catchments B11 and B12 fall at the limit of the desert and stepped regions, even according to the classification of Ponce et al. [121], these basins correspond to the semi-arid zone, which is related to complex processes such as a base flow essentially absent, prolonged wet or dry sequences, and rainfall that tends to be more variable in both space and time than in humid regions [120].

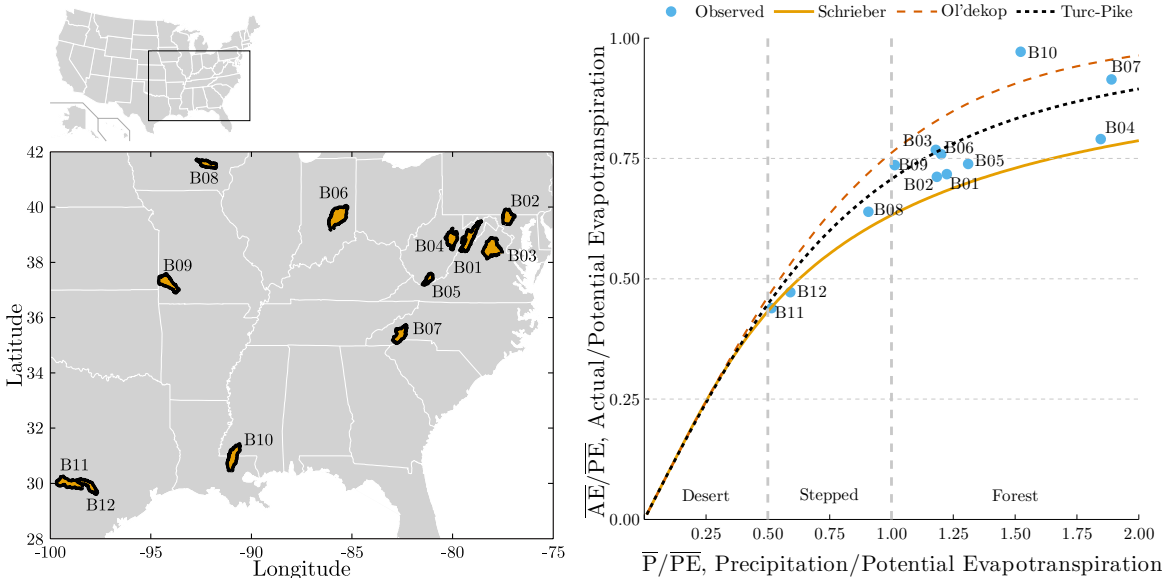
The hourly precipitation datasets were developed by the National Weather Service (NWS)-Hydrology Laboratory based on hourly and daily rain gauge data gathered from the National Climate Data Center (NCDC). The daily streamflow datasets were obtained from the US Geological Survey (USGS). The climatic potential evaporation data was derived from the National Oceanic and Atmospheric Administration (NOAA) Freewater Evaporation Atlas [59].

[†]The database can be found at `http://www.nws.noaa.gov/oh/mopex/mo_datasets.htm`.

Table 6.4: Main characteristics of the studied catchments.

Basin code		Area	Elev.	Soil	Veg.	Precipitation		Streamflow		P/PE
USGS	Here	(km^2)	(m)	type	type	$\overline{P}$ (mm)	CV	$\overline{Q}$ (mm)	CV	
01608500	B01	3810	171	L	DB	2.55	0.13	1.06	0.31	1.22
01643000	B02	2116	71	SL	DB	2.91	0.16	1.16	0.36	1.18
01668000	B03	4134	17	CL	MF	2.97	0.14	1.04	0.36	1.18
03054500	B04	2372	390	L	DB	3.60	0.13	2.06	0.21	1.85
03179000	B05	1020	465	SCL/L	DB	2.66	0.12	1.16	0.28	1.31
03364000	B06	4421	184	SL/CL	CL	2.82	0.13	1.04	0.31	1.20
03451500	B07	2448	594	L	MF	4.24	0.14	2.19	0.25	1.89
05455500	B08	1484	193	CL	CL	2.47	0.21	0.73	0.65	0.91
07186000	B09	3015	254	SL/CL	DB	3.04	0.17	0.83	0.59	1.01
07378500	B10	3315	0	SL	EN	4.48	0.15	1.62	0.34	1.52
08167500	B11	3406	289	C	CL/NV	2.16	0.32	0.32	0.90	0.52
08172000	B12	2170	98	C	CL/NV	2.34	0.29	0.47	0.66	0.59

Elev.: elevation above mean sea level, Veg.: vegetation, CV: coefficient of variation, L: loam, S: silt, C: clay, DB: deciduous broad leaf, MF: mixed forest, CL: croplands, EN: evergreen needleaf, NV: native vegetation. Mean annual values (from 1960 to 1998).



(a) Basins location

(b) Empirical curves for evaporation

Figure 6.4: Basins and hydroclimatological regimes.

Daily minimum and maximum temperatures datasets were obtained from National Center for Environmental Predictions/National Center for Atmospheric Research (NCEP/NCAR) Global Reanalysis data [87].

6.4 Results and discussion

6.4.1 Performance of the baseline model (R100P)

To show that the strength of the baseline model (R100P) is not based on the efficiency of a particular model, in this case the FFNN, but in the conceptual aspects behind HEPS such as diversity or complementarity between predictors, we reference the baseline results with respect to several ANN architectures explored by Vos [149] for the same basins evaluated here (Fig. 6.5).

It is important to note that our definition of the training and test subsets corresponds with the experimental design of Vos [149], this coincidence emerges from the experimental conditions established in the paper that summarizes the second and third workshop of the MOPEX project [53], which presents the details of the database and procedures to allow comparisons between results obtained by different research groups.

In Fig. 6.5, presented by Vos [149] for comparison of various ANN architectures, except for the results of the baseline model (R100P) indicated by the dotted horizontal line, he evaluated a Persistence model (PM), a Multiple linear regression model (LIN), two variants of FFNN, two variants of Elman recurrent ANN (EL), four variants of Williams-Zipser fully recurrent ANN (WZ), ESN and two of its variants, which are known for their high performance. More details about this particular set-up can be found in Vos [149].

In general, Fig. 6.5 shows a high dispersion of FFNN results compared to other ANN structures. Also, the median of the results confirms that the other ANNs are generally better than the FFNN structure. However, the results of the R100P model, which integrates an ensemble approach, reveal three main features:

1. The baseline model presents a substantial gain with respect to the results of any single FFNN.
2. In most cases, the baseline model equals or improves results of recurrent architectures, except in basins B01, B03, B09, and B12 where some recurrent structure presents slightly higher performances.
3. The results of the baseline model are not very far from the best results, that are generally obtained by ESN variants, except for basins B01, B02, B09, and B11 where such models are substantially better. However, in three basins (B03, B06, and B07) results from the baseline model outperform all other ANN architectures.

It is worth noting that our interest, under the ensemble philosophy, is part of a more cooperative than competitive framework between different prediction models, therefore we can expect a system with higher performance combining efficiently the different types of ANN structures exploiting the strengths of each architecture.

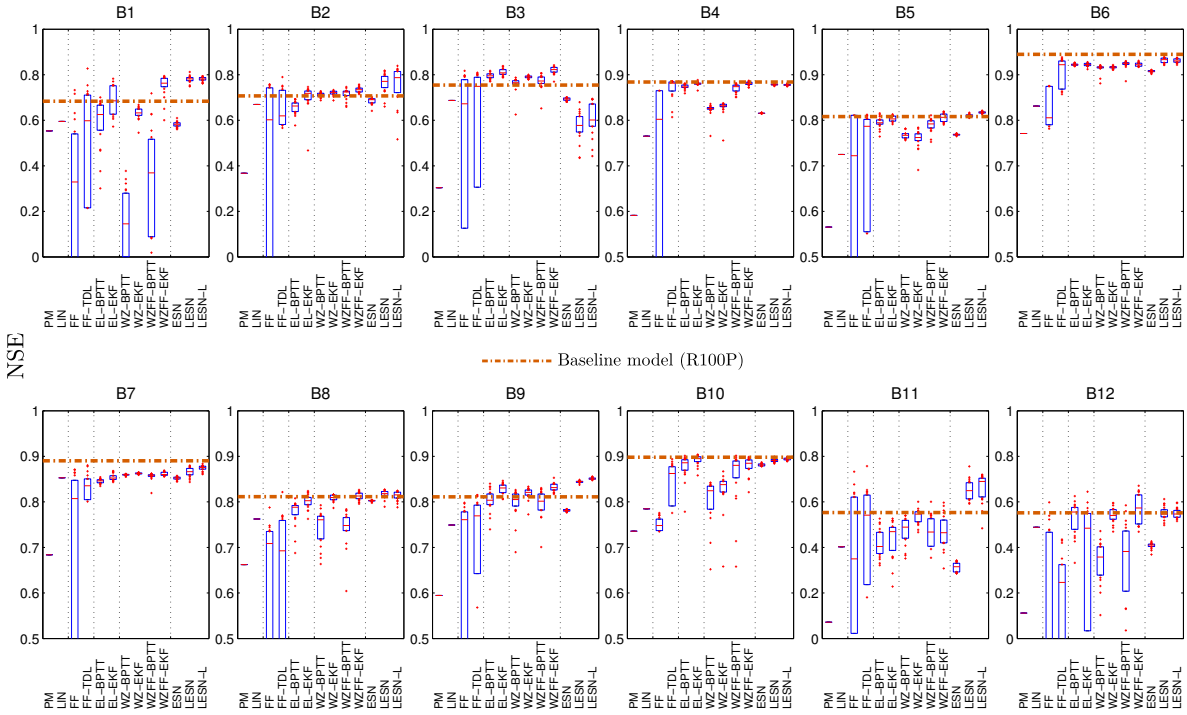


Figure 6.5: Test performance of several ANN architectures and the R100P model in terms of NSE.The box plot expresses statistics of over 20 models runs. The central mark is the median, the edges of the box are the first and third quartiles, and points outside that range are plotted individually as dots. Note that not all subfigures have the same scale.

6.4.2 Comparison of training and test properties

Figure 6.6 presents an exploratory data analysis regarding the train and test datasets for precipitation, streamflow, and aridity index. Ideally, the average values in the train and test datasets should be equal. Fig. 6.6a, b, and c show a good approximation to this condition given that the mean values are grouped around the the diagonal. However in Fig. 6.6b, basins B01, B02, B03, B4, B09, and B10 present training values slightly higher than the test dataset, which is not a problem in ANN modelling. In contrast, basin B07 presents the "non-ideal" condition in the ANN experiment design, given the limited capabilities of the ANN extrapolation [80]. Nevertheless, in the case of basin B07, this small difference does not greatly reflect on the PI illustrated in Fig. 6.6d.

Figure 6.6d illustrates the negative relation between the coefficient of variation of the maximum

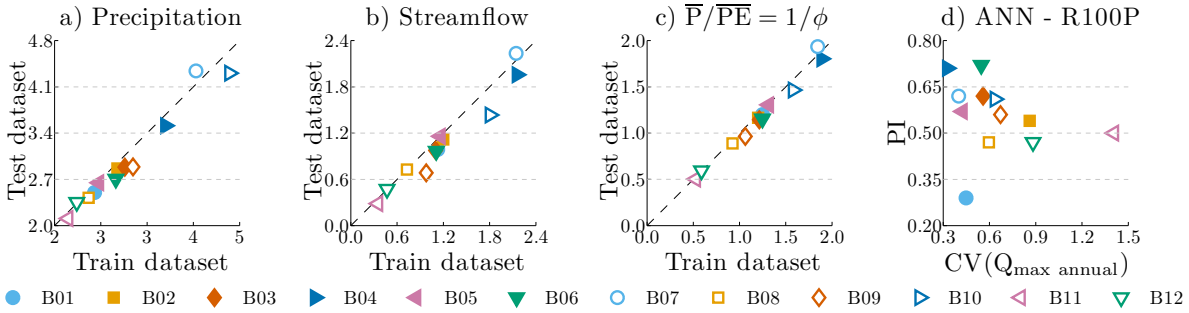


Figure 6.6: Train and test datasets properties.

annual streamflow series, CV($Q_{max\ annual}$), and the PI, except for basins B01 and B10. Basins B02, B12, and even further basin B11 present high coefficients of variation which could result from the stratification schemes, as it will be discussed in the next section. Note that the CV($Q_{max\ annual}$) measures roughly the irregularity or complexity of the series. The poorer performance of the ANN of reference (R100P model) in basins B11 and B12 stresses the difficulty of forecasting in semiarid zones.

The coefficient of variation of the mean annual precipitation in Table 6.4 reflects the high variability of basins B11 and B12 with respect to the others. In general, the processes in such semi-arid zones are hardly replicable in hydrological models as in any regression model, because the model parameters may differ in prolonged wet or dry periods [120].

6.4.3 Stratification results

Results detailed in this section correspond to the test subset, which contains examples not used in ANN training. Figure 6.7 shows the average results of 30 experiments with ANN stacks trained with four resampling schemes. The first is based on a random selection of data and the other correspond to stratified resampling schemes with respect to the Input space (I), the Output space (O) or both the Input and Output spaces (I+O). Each scheme is evaluated in four scenarios depending of the number of data used in training (12.5%, 25%, 50%, and 100%). Each point represents the MSE gain of each scheme in relation to the reference scheme R100P (Eq. 6.3). Figure 6.7 exhibits the relationship between the expected error and the availability of data to train. So, the comparison of the error obtained between two incremental scenarios generally shows a negative trend in terms of gain, except in some cases where the trend is positive, see for example the $50\% - 100\%$ section of basin B01 or the $12.5\% - 25\%$ section in some schemes of basins B02, B03, B04, and B11.

In addition, it can be seen that in 88% of cases, the stratification schemes improve the stacked ANN performance compared to a random data selection. Despite this, it is important to highlight that, except for basin B01 and the stratification schemes with a 100% of available

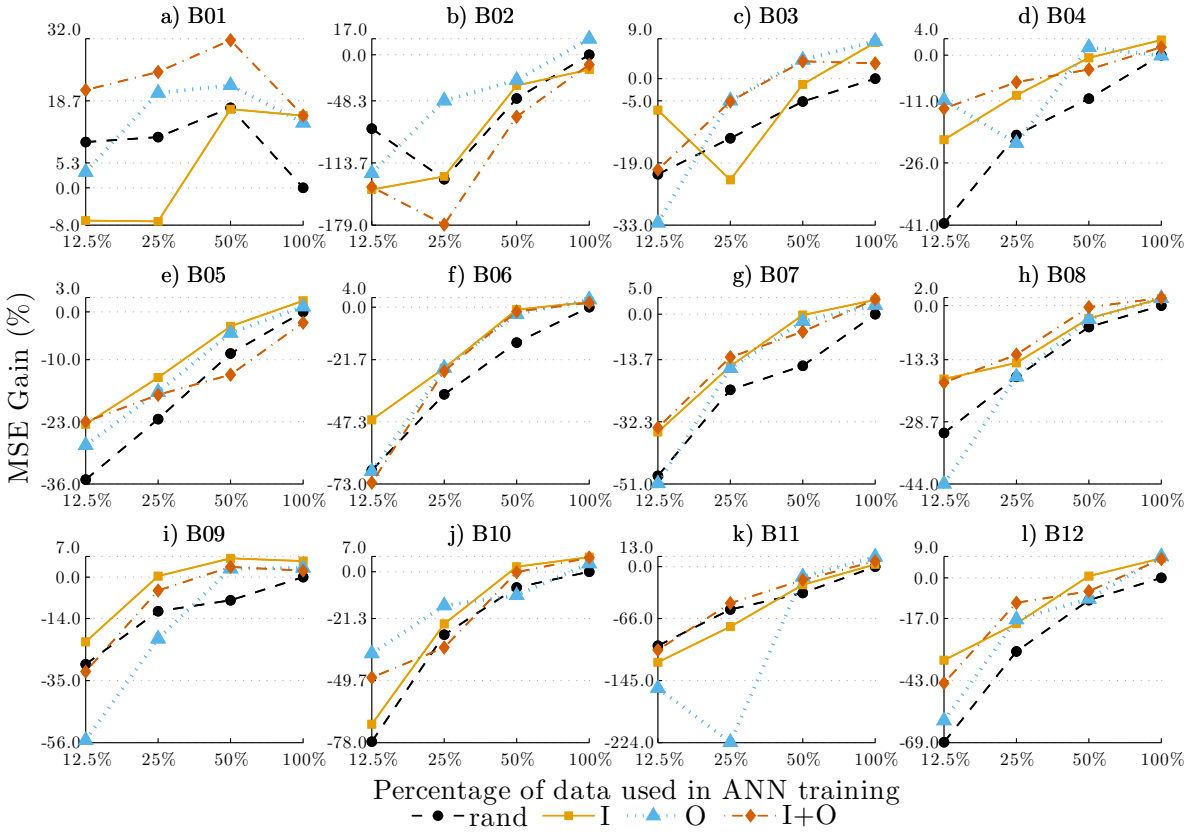


Figure 6.7: Stratification results. I: input scheme. O: Output scheme. I+O: Input and Output scheme. Note that not all subfigures have the same scale.

training information, the gain is negative. With respect to determining the best stratification scheme, the specific characteristics of each basin and the availability of data should be considered. A summary of results of stratification is presented in Table 6.5, which shows the best schemes for each basin.

We adopt the median as a measure of central tendency given the high dispersion of results, principally for basins B01, B02, and B11, which present the best gains when the stratification is performed on 100% of the training data. It highlights the poor performance of the 12.5% scenario, which shows that, in basins B02 and B11, the random selection is the best alternative. Regarding the 25% scenario, we have a median loss of 12.5% that can be considered acceptable in repetitive tasks such as determining the optimum configuration of the ANN or the input variable selection.

The median results for the 50% scenario shows that the stratified schemes lead to models with equal performance to the R100P model but using half the data. With regard to the 100% scenario, the median indicates that the model performance can be improved by about 6%.

Table 6.5: Mean MSE gain of resampling techniques relative to R100P scheme (see Eq. 6.3). Analyzed schemes: random (R), stratified input space (I), stratified output space (O) and stratified input and output spaces (I+O). Md and Mo(BS) in last row represent the median and the most frequent best scheme (statistical mode), respectively.

Basin code	MSE R100P	Percentage of data used in ANN stack training · Best Scheme							
		12.50%	BS	25.00%	BS	50.00%	BS	100.00%	BS
B01	0.897	20.97	I+O	24.85	I+O	31.69	I+O	15.49	I
B02	1.751	−77.66	R	−48.15	O	−26.65	O	16.79	O
B03	0.731	−7.10	I	−4.99	I+O	4.24	O	8.43	O
B04	0.977	−10.85	O	−6.51	I+O	1.92	O	3.68	I
B05	0.780	−22.99	I+O	−13.75	I	−3.02	I	2.30	I
B06	0.162	−46.53	I	−25.18	I	−1.01	I	3.14	I
B07	0.408	−34.12	I+O	−12.95	I+O	−0.29	I	4.46	I
B08	0.506	−18.14	I	−12.08	I+O	−0.40	I+O	1.89	I+O
B09	0.570	−21.91	I	0.31	I	6.33	I	5.42	I
B10	0.906	−37.36	O	−15.42	O	2.23	I	6.64	I
B11	0.445	−100.84	R	−46.55	I+O	−13.01	O	12.28	O
B12	0.372	−34.47	I	−10.46	I+O	0.73	I	8.89	O
Md · Mo(BS)		−28.55	I	−12.51	I+O	−0.22	I	6.03	I

In order to define the relevance of stratification schemes, the statistical mode or most frequent value is displayed next to the median value. Thus, for a stratification of 25% of the information, the best model is the one based on the input and output space (I+O). Similarly, the scenarios corresponding to 50% and 100% generally present better a performance when the stratification is performed only in the input space (I).

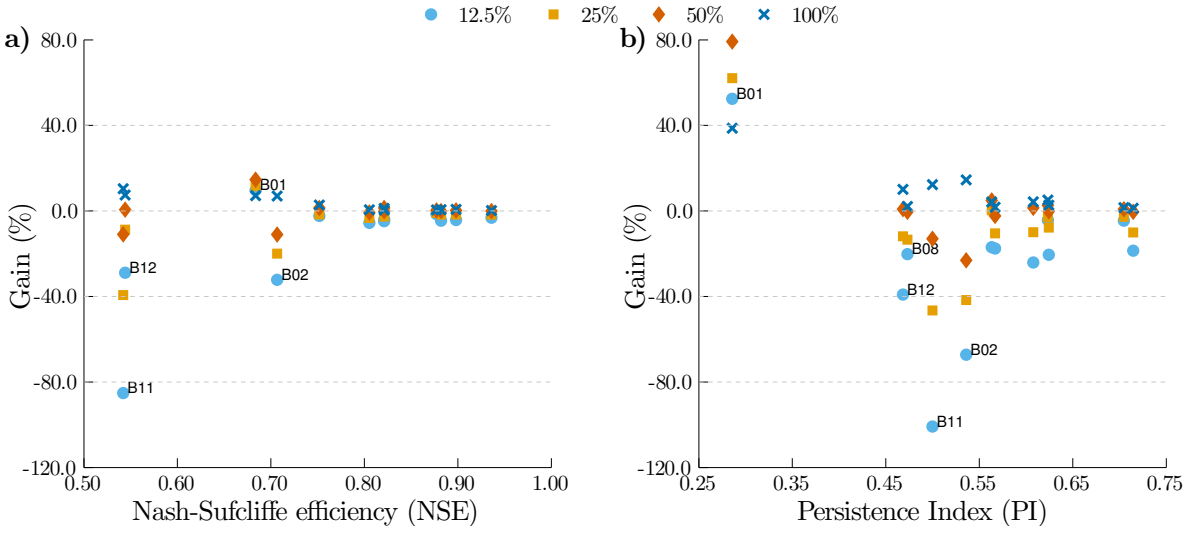


Figure 6.8: MSE normalizations and best individual stratification schemes.

Figure 6.8 gathers the results of the schemes shown in Table 6.5 with respect to two normalizations of the MSE, the NSE, and the PI, defined in Sect. 6.2.3. Regarding the NSE, we can appreciate the following:

- The results of the basins with NSE higher than 0.75 (Fig. 6.8a) are concentrated in the range between 6% and −3%. This is particularly relevant given that in the scenario with 12.5% of data, the MSE gain showed a median loss equal to −28.55% (Table 6.5).
- Basins with the smaller initial performance, i.e. basins B01, B02, B11, and B12 with (NSE < 0.75), present high variability in stratification schemes, but at the same time, present the better gain using stratification with 100% of the information. Note the negative relationship between the NSE and gain in the 100% scenario (cross markers).
- The atypical behaviour of basin B01 is again ratified in the contradictory relationship between the gain and percentage of data used to ANN stack training.

With respect to the PI, Fig. 6.8b shows that this skill score is more difficult to preserve than the NSE, in a very similar trend to MSE, because the median values for the scenarios 12.5%, 25%, 50%, and 100% are −19%, −10%, 0.3%, and 4% respectively. As expected, basins B01, B02, B11, and B12 follow the same trend of high variability.

6.5 Conclusion

First, it should be noted the good deterministic performance of the proposed ensemble model compared to other ANN architectures such as partially and fully recurrent ANN and the novel technique called ESN. At this point, it is important to focus on the basic principle of ensemble methods: complementarity between models reduce the systematic bias of the system.

The relevance of the stratification schemes was evaluated with respect to an ANN configuration based on a random selection of data for estimation and validation of network weights (training), using the early stopping method and an 18-year dataset for training (R100P). Note that, for all results presented in this chapter, we used a testing dataset formed with data never used in the ANN training stack phase.

It is generally accepted that the hydrological processes of semiarid basins (B11 and B12), defined by the aridity index, are more difficult to simulate, as confirmed by the lower performance obtained from the R100P reference model as well as from the proposed stratification models. Other ANNs may behave better in such settings, as shown by Vos [149] in a study conducted with the same database analyzed here.

Concerning the data stratification used in ANN training, we tested the methodology developed by López et al. [96], for which data resampling is based on the evaluation of the k-means clustering algorithm and distribution of data within each cluster in folds according to the distance to the cluster centre. Results showed the importance of the input space, confirming

the findings of Abrahart and See [2] and Anctil and Lauzon [8], which resorted to Kohonen maps clustering. It is worth mentioning that the Machine Learning community most commonly exploits stratification in the output space [51]. Here, the results showed that a pertinent selection of 25% or 50% of the available training information made possible achieving test performances similar to the reference model and, in some cases, improving on it despite a lower number of training data.

We conclude that despite the diversity imposed at the data level, in some cases, it was not detrimental to the high model performance evaluated in the R100P scheme. We found that the best scheme used 50% of the training information, selected from the input space only. Reducing further the size of the training set had a negative impact on the performance. However, it would be interesting to evaluate how much to compromise deterministic performance depending on the diversity and reliability of the ensemble, because we obtained average loss of only 12.5% in the 25%-stratification scheme based on the input and output spaces.

Finally, based on the results of this chapter and the concepts behind the Multi-Level Diversity (MLD) model, we propose in the next chapter the integration of another source of diversity at the input model level. In the further pursuit of diversity, future work should integrate structural level evaluation from a system composed of multiple ANN architectures.

Chapter 7

Diversity from Dataset, Parametric, and Model Inputs Levels

In this chapter, we propose a framework based on two separate but complementary topics: data stratification and Input Variable Selection (IVS). We promote an Artificial Neural Network (ANN) prediction ensemble in which each predictor is trained based on input spaces defined by the IVS application on different stratified sub-samples. All this, added to the inherent variability of classical ANN training with gradient methods, leads us to our ultimate goal: diversity in the prediction, defined as the complementarity of the individual predictors.

The stratification, evaluated in Chap. 6, showed that the informativeness of the data is far more important than the quantity used for ANN training. Here, exploiting the same database than for the last chapter, we show that ensembles designed from ANNs trained on different sets of input variables and 50% of the available data lead to efficient probabilistic models. Skill evaluation is again based on a Ensemble of 30 FFNNs trained with early stopping using a **R**andom sampling of **100 P**ercent of the available information and a single predefined set of inputs variables (R100P model).

Results show that from a deterministic view, the main advantage is the efficient selection of the training information, which is an equally important concept for the calibration of conceptual hydrological models. On the other hand, the diversity achieved is reflected in a substantial improvement in the scores that define the probabilistic quality of the Hydrological Ensemble Prediction Systems (HEPSs).

7.1 Introduction

Probabilistic forecast can be associated with multi-model or ensemble forecast approaches, from a non-reductionist viewpoint. The modeller is interested not only in finding the best prediction but also in obtaining the best estimate of the forecasts uncertainty [21]. Conse-

quently, the modeller uses the statistical tools described in Sect. 1.3 to evaluate the quality of ensemble predictions. In streamflow probabilistic forecasting, the *cascading model uncertainty* [115] evaluates the uncertainty from different sources: climatological variables, structural conceptualization, and parametric variability. The justification of this level of complexity is the potential economic value in a decision making scenario [127], which should be evaluated after ensuring the quality of the system [111].

Hydrological prediction scenarios emerged as a result of the acceptance that no particular model in existence today is superior to other models for all type of applications and under all conditions [54]. In the machine learning community, the previous sentence is known as the "No Free Lunch" theorem [7, 47], which states that no algorithm may be assumed to be better than any other algorithm when averaged over all possible types of problems. Consequently, the multi-model approach is a common area in both the hydrological and machine learning communities. The classical evaluation of multiple forecast scenarios leads to the use of reductionist decision schemes based on combining functions such as average or weighted combination, ignoring the importance of reliability, resolution, sharpness, and consistency of the set of scenarios as an integral part of the prediction system (see Sect. 1.2).

At this point, the evaluation of an ANN stack or model aggregation in which each predictor represents different ANN initializations and therefore different parameters of the same structure, stands out as an efficient way to reduce the bias of prediction [8, 132, 163]. However, several authors [21, 27, 28, 29] showed that the gain in bias does not follow the same trend with respect to the reliability of the prediction system, which reduces its quality from the probabilistic viewpoint, because this property is strongly related to the consistency and prediction system value [111].

On the other hand, other authors [36, 92] have shown, from a deterministic point of view, that the success of an ensemble prediction mainly lies in the diversity of the ensemble, defined as the complementarity of the individual predictors. Intuitively, we want the ensemble members to be as correct as possible, and in case they make errors, these errors should be on different data-points. However, the diversity of the ensemble itself is not enough, we need to get the right balance between diversity and individual accuracy, in order to achieve the lowest overall ensemble error [36]. Also, the quantification of diversity represents another problem; although several measures have been proposed [92], they do not reveal how to achieve diversity [36]. The efficiency shown by methods such as Adaptive Boosting (AdaBoost) [63] and its variants [136] is remarkable since they use the diversity explicitly during the process of ensemble building.

Our hypothesis is that active manipulation of diversity and the accuracy of each predictor (or stack member) not only decreases the system bias but indirectly improves the system reliability. To test this, we use partially the MLD model (Sect. I.4) proposed by Kuncheva [92], which generalizes the levels of modelling uncertainty at four levels: variability in data

subsets (uncertainty from data), manipulation of different inputs subsets (uncertainty from inputs), testing diverse models and/or different parameter settings (uncertainty from structural parametrization of models), and different techniques for fusion or predictors selection (uncertainty from combiner functions).

The HEPS complexity, related to the number of members to manipulate at each time step, is evident. For example, Velázquez et al. [148] presents a system of 800 members combining probabilistic meteorological information from more than a dozen hydrological models (see Chap. 3, 4, and 5). Similarly, machine learning offers numerous techniques such as support vector machines, ANN, decision trees, and fuzzy logic, among others, that are also accompanied by the typical parametric uncertainty. For example, Caruana et al. [41] show a system of 2000 scenarios in the context of handwritten character recognition.

However, in this chapter, we propose a stack of ANNs for which each member represents different problem domains and ensures their ensemble accuracy without the need to evaluate hundreds of scenarios, arbitrarily limiting the stack to thirty predictors. Note that another way to optimize the prediction system is based on the concept of "overproduce and select", generating a pool of predictors to later select those that together optimize an error function. In Chap. 3, 4, and 5, we presented a detailed overview of this approach in the ensemble prediction context with several hydrological lumped models.

In this chapter, we explore different techniques that lead to the construction of a HEPS based on ANN, and the concept of diversity addressed from three levels of variability: selection of data, inputs subsets, and parametric uncertainty. For this, we tested our hypothesis on the database used in Chap. 6, i.e. the twelve basins from the MOPEX project [53].

Because our focus is also on the ensemble variability due to different configurations of the input space, we explore a framework in which each member of the ANN ensemble is trained using certain data and occasionally a particular configuration of the input space, propagating data-level variability into the input space. So, we encourage dynamic selection of inputs with the FGS method [9, 10, 13]. In this regard, several authors have highlighted filter-type selection methods based on mutual information [24, 37, 104] arguing, in many cases, the high computational cost of FGS. However, in our case of forecasting at daily resolution, such computational cost is not a constraint; simplicity is more attractive.

Note that the parametric variability of each ANN is evident since we use different initializations and different data to train them, also the optimization algorithm is based on a local search procedure. It is worth noting that diversity may also focus on the structure of each ANN, as suggested by Brochero et al. [30], or in the level of selection or combination of predictors. However, in this context, such sources of variability are designated as subject of future work.

It is important to emphasize that our goal moves away from the "perfect" model paradigm

embracing a probabilistic formulation. Coinciding with one of the guidelines outlined by Abrahart et al. [6] regarding the directions that should be taken in an operational ANN integration goal, we thus propose the probabilistic evaluation of the forecasts. We outline the methodology in Sect. 7.2, results and discussion are presented in Sect. 7.3, finally, in Sect. 7.4, conclusions are drawn and a guideline for future work is given.

7.2 Methodology

We propose the evaluation of two independent but complementary topics such as data stratification (Sect. 6.2.1) and IVS on the 12 basins described in Sect. 6.3. The link between both topics is the active pursuit of diversity.

The main idea of this chapter is to encourage diversity in an ensemble of 30 ANNs based on three levels of uncertainty outlined in the MLD model: parametric variability, the selection of data to train the model, and the selection of the model inputs.

The parametric variability of each member is the product of random ANN weights initialization, in conjunction with the Levenberg Marquardt optimization algorithm, which relies on a search-based local gradient.

Regarding the selection of the data, in the preceding chapter it was shown that if each member is trained with a stratified selection of 50% of the data, it is possible to diversify the ensemble and obtain good results from the average prediction. In summary, the stratification methodology allows suitably and systematically choosing sub-samples representing different conditions for which the ANN should react. Importantly, the stratification results are variable because the methodology generates multiple stratified sets and the selection of one of them is random. Additionally, the clustering algorithm used (k-means), which is known for its stability, did not guarantee a global optimum convergence.

With respect to the selection of inputs, in this chapter we couple the stratification concept with the IVS problem to promote ensemble diversity. In this case, we propose that each ensemble member represents the results of a stepwise IVS incorporating the stratified data selection. Such a procedure, presented below, will be called hereafter Dynamic Input Spaces imposed by Stratified Examples propagated on artificial Neural networks Training (DISSENT). Importantly, this mechanism is considered dynamic in the sense that the ensemble is not trained with a single set of input variables.

7.2.1 DISSENT method

Figure 7.1 shows the flowchart of the DISSENT method, which is applied 30 times in order to setup the ANN ensemble of 30 members. This process is described in the following steps:

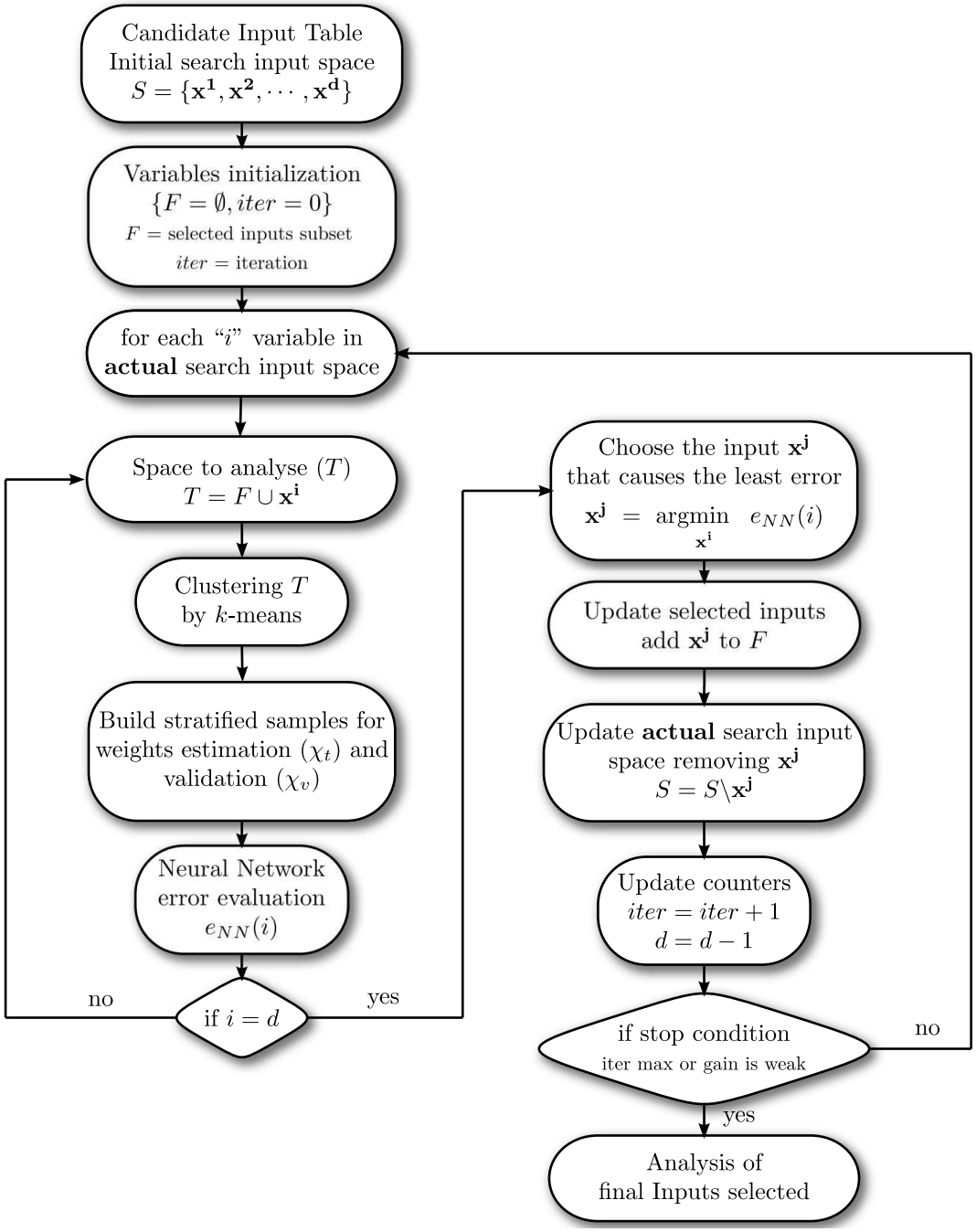


Figure 7.1: Dynamic IVS procedure.

Step 1: Global parameters set-up. Based on the results discussed in Sect. 6.4.3 and according to the approach of Abrahart and See [2], and Anctil and Lauzon [8], this phase focuses only on the evaluation of the stratification in the input space; however, only a few minor adjustments are required to extend the methodology to other methods of stratification. In this way, it is only necessary to define the percentage of data to generate the stratified sub-sample (p) and the percentage of this sub-sample to be used in the estimation dataset

configuration (p_{est}), 75% in our case – the remaining 25% is used in the validation phase of the early-stopping mechanism.

Step 2: Candidate input table. The successful development of ANN models depends largely on the availability of pertinent model inputs. In the present study, we use a candidate input table of twenty-five potential variables that represent climatological and streamflow lagged series, which provides dynamic information to the hydrological process (Table 7.1). So, for each member in the ensemble, we apply an IVS based on the candidates variables. We must define the best five inputs in order to provide a fair comparison with the baseline model presented in Sect. 6.4.1, in which inputs correspond to variables highlighted in bold in Table 7.1.

Table 7.1: List of model input candidates.

#	Input	Description
1	Q_{t-1}	**Previous-day streamflow**
2	Q_{t-2}	**Previous-2-day streamflow**
3	Q_{t-3}	Previous-3-day streamflow
4	P_{t-1}	**Previous-day precipitation**
5	P_{t-2}	**Previous-2-day precipitation**
6	P_{t-3}	**Previous-3-day precipitation**
7	ET_{t-1}	Previous-day evapotranspiration
8	ET_{t-2}	Previous-2-day evapotranspiration
9	ET_{t-3}	Previous-3-day evapotranspiration
10	Tmx_{t-1}	Previous-day maximum temperature
11	Tmx_{t-2}	Previous-2-day maximum temperature
12	Tmx_{t-3}	Previous-3-day maximum temperature
13	Tmn_{t-1}	Previous-day minimum temperature
14	Tmn_{t-2}	Previous-2-day minimum temperature
15	Tmn_{t-3}	Previous-3-day minimum temperature
16	ΔQ_{t-1}	Previous-day historical streamflow increment: $Q_{t-1} - Q_{t-2}$
17	ΔQ_{t-2}	Previous-2-day historical streamflow increment:$Q_{t-2} - Q_{t-3}$
18	ΔP_{t-1}	Previous-day historical precipitation increment:$P_{t-1} - P_{t-2}$
19	ΔP_{t-2}	Previous-2-day historical precipitation increment:$P_{t-2} - P_{t-3}$
20	ΔET_{t-1}	Previous-day historical evapotranspiration increment:$ET_{t-1} - ET_{t-2}$
21	ΔET_{t-2}	Previous-2-day historical evapotranspiration increment:$ET_{t-2} - ET_{t-3}$
22	ΔTmx_{t-1}	Previous-day historical maximum temperature increment:$Tmx_{t-1} - Tmx_{t-2}$
23	ΔTmx_{t-2}	Previous-2-day historical maximum temperature increment:$Tmx_{t-2} - Tmx_{t-3}$
24	ΔTmn_{t-1}	Previous-day historical minimum temperature increment:$Tmn_{t-1} - Tmn_{t-2}$
25	ΔTmn_{t-2}	Previous-2-day historical minimum temperature increment:$Tmn_{t-2} - Tmn_{t-3}$

Step 3: Variables initialization. We appoint an empty set as initialization of the selected variables (F) and the set of variables to evaluate (S) equal to the pool of input candidates (Table 7.1). The number of variables to be evaluated in each iteration is initially equal to the number of input candidates, 25 in this case. This value decreases proportionally with the iteration ($iter$), which is initialized to zero, so $d = 25 - iter$ at each iteration.

Step 4: Local analysis of variables. For each variable of the search space, the stratification is evaluated in the union of the selected variables (F) and the variable in analysis. In the first iteration, this process is performed on each of the input candidates since F is empty. Subsequent iterations consider the selected variable in the past iteration.

Once the stratified estimation and validation subsets are configured, the process continues with the evaluation of the ANN described in Sect. 6.2.2, for later storage of its performance as a function of the input evaluated. Although the validation set is used to avoid overfitting in an early-stopping or cross-validation tasks, the training error is calculated using both estimation and validation subsets, as suggested by Haykin [77]. Note that given the instability of the ANNs, results for each iteration represent the best ANN between 10 different initializations based on the MSE minimization.

Step 5: Variable selection. After repeating last step on all candidate variables, we choose the variable that contributes most to the decrease of MSE. Then, the best variable of each iteration is stored in F and removed from the search space S to be used in the next iteration. Additionally, counters *iter* and d are updated.

Step 6: Stop condition. For a direct comparison of the DISSENT model with the five predefined variables evaluated in Sect. 6.2.1, the selection process stops at the fifth iteration. In an optimization context, we recommend a more rigorous application based on a threshold gain to accept the inclusion of a new variable, as proposed by Anctil et al. [13], who uses a gain threshold of 10%, adopting the same procedure proposed by Senbeta et al. [133] in the development of conceptual hydrological models.

Figure 7.2 is an example of the dynamics of the DISSENT procedure. Here, we show the mechanism for configuring each of the ensemble members. In the first iteration, we evaluate the performance of each of the 25 candidate variables, extracting, in a stratified fashion, 50% of the data to determine the performance of a single ANN. Remember that this performance includes the training and validation error, it is based on the best result of 10 ANN evaluations differentiated by weights initialization.

Assuming that X_4 is the variable that minimized the error the most, in the second iteration, the stratification is evaluated in the conjunction of X_4 with other individual variables to determine the second most influential variable in the minimization of the error, in this case for example the variable X_1. The process continues until the 5 best variables that define one of the members of the ensemble are identified.

Note that in essence, the systematic selection of 50% of the data, which has a random component in the final selection of the stratified sets, the local nature of greedy IVS, as employed here, and the ANN parametric variability, depending on the initial conditions of the optimization, lead to a diversified ensemble conception.

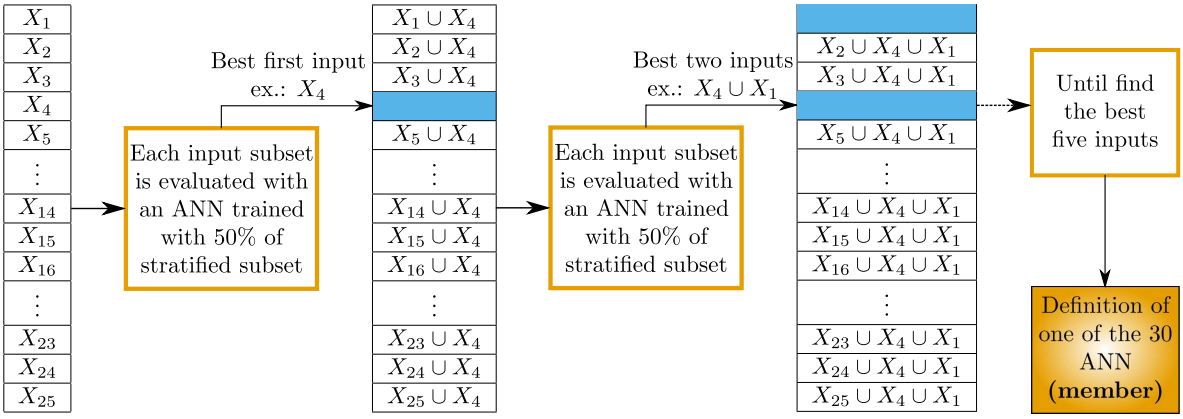


Figure 7.2: DISSENT procedure example.

7.2.2 Performance evaluation

Relevance of the DISSENT procedure is analyzed again based on the performance of an Ensemble of 30 FFNNs trained with early stopping using a **R**andom sampling of **100 P**ercent of the available information and a single predefined set of inputs variables (R100P). Unlike the precedent chapter, which is focused on deterministic performance functions, in this chapter we also evaluate the probabilistic performance of the ANN ensemble. Consequently, we employ the mean CRPS, the mean IGNS, the error in the reliability diagram (RD_{MSE}), and the normalized deviation of the rank histogram from flatness (δ ratio). These measures, negatively oriented, are described in detail in Sect. 1.3. Except the δ ratio, a kernel density estimation of the prediction ensembles is used in scores evaluation.

To facilitate an analysis independent of the basins scale and to obtain a robust performance estimator, we propose a gain index based on the median error of 30 evaluations of the DISSENT model ($me_{DISSENT}$) and the median error of 30 evaluations of the R100P model (me_{R100P}):

$$G = \frac{me_{R100P} - me_{DISSENT}}{|me_{R100P}|}. \tag{7.1}$$

A positive index indicates superior performance of the DISSENT model. The absolute value in the denominator is needed to assess the performance of the IGNS, which can have positive and negative values. The median error (me) is used as measure of central tendency due to the asymmetry of errors, especially in the random cases.

7.3 Results and discussion

7.3.1 Inputs sets

Table 7.2 shows the number of input subspaces that were identified after conducting 30 times the evaluation of the methodology shown in Fig. 7.1. It is noteworthy that, except for basin

B02, which shows great variability in the selection of variables, all the basins show five different schemes for the input space. In Table 7.2 #IS indicates the number of different inputs subspaces found. Best input schemes shows the best variables according to the mean rank of selection. Note that the variables order in Table 7.2 is related to their order of importance.

Table 7.2: Number of input subspaces found in 30 DISSENT experiments.

Basin code	#Input schemes	Best input scheme				
B01	5	Q_{t-1}	P_{t-1}	ΔQ_{t-1}	ΔP_{t-1}	P_{t-3}
B02	10	Q_{t-1}	P_{t-1}	ΔP_{t-1}	ΔQ_{t-1}	ΔTmx_{t-1}
B03	5	Q_{t-1}	P_{t-1}	ΔP_{t-1}	ΔQ_{t-1}	ΔET_{t-1}
B04	5	P_{t-1}	ΔQ_{t-1}	Q_{t-1}	Q_{t-2}	Tmx_{t-3}
B05	5	Q_{t-1}	P_{t-1}	Q_{t-2}	Tmx_{t-3}	ΔTmx_{t-2}
B06	5	Q_{t-1}	P_{t-1}	ΔP_{t-1}	P_{t-2}	Q_{t-2}
B07	5	Q_{t-1}	P_{t-1}	Q_{t-2}	Q_{t-3}	ET_{t-2}
B08	5	Q_{t-1}	P_{t-1}	Q_{t-2}	Tmx_{t-2}	ΔP_{t-1}
B09	5	P_{t-1}	Q_{t-1}	Tmn_{t-3}	ΔQ_{t-1}	Q_{t-2}
B10	5	ΔQ_{t-1}	Q_{t-1}	Q_{t-2}	P_{t-1}	ΔP_{t-1}
B11	5	P_{t-1}	ΔQ_{t-1}	Q_{t-3}	Q_{t-1}	Q_{t-2}
B12	5	Q_{t-1}	P_{t-1}	ΔTmn_{t-1}	Q_{t-2}	ΔQ_{t-1}

To get an idea of the overall importance of the analyzed variables, we estimate the best IVS scheme based on the mean selection rank of each variable within the DISSENT process. That is, if the variable Q_{t-1} is chosen as the best variable in the first iteration, its rank is equal to one. In the following four iterations, we rank the other four variables, the rest of the variables will be penalized by the maximum possible rank, i.e. 25, coinciding with the number of input candidates (Table 7.1). Finally, the mean selection rank of each variable is calculated based on the 30 experiments and then, we chose the five variables with the lowest ranks. Another possibility, in a competitive framework, would be to choose the best set of variables based on the minimization of an error function.

As expected, variables with the greatest participation are streamflow and precipitation. Nevertheless, with the aim to promote diversity, we consider that all schemes are equally important. As an example, Fig. 7.3 presents the relative histogram of the selection found for basins B02, B04, B06, and B11. As it can be seen, the minor differences between the input subsets are centred on variables such as temperature or evapotranspiration that do not commonly participate to daily forecasting models. In general, it is observed that the streamflow and precipitation of the previous day, as well as their increments, are more relevant. Nonetheless, basin B11 exhibits a particular dependence of streamflow of the previous 3 days and the maximum temperature of the previous day, which can be related to the complexity of the timing between different variables in semiarid basins.

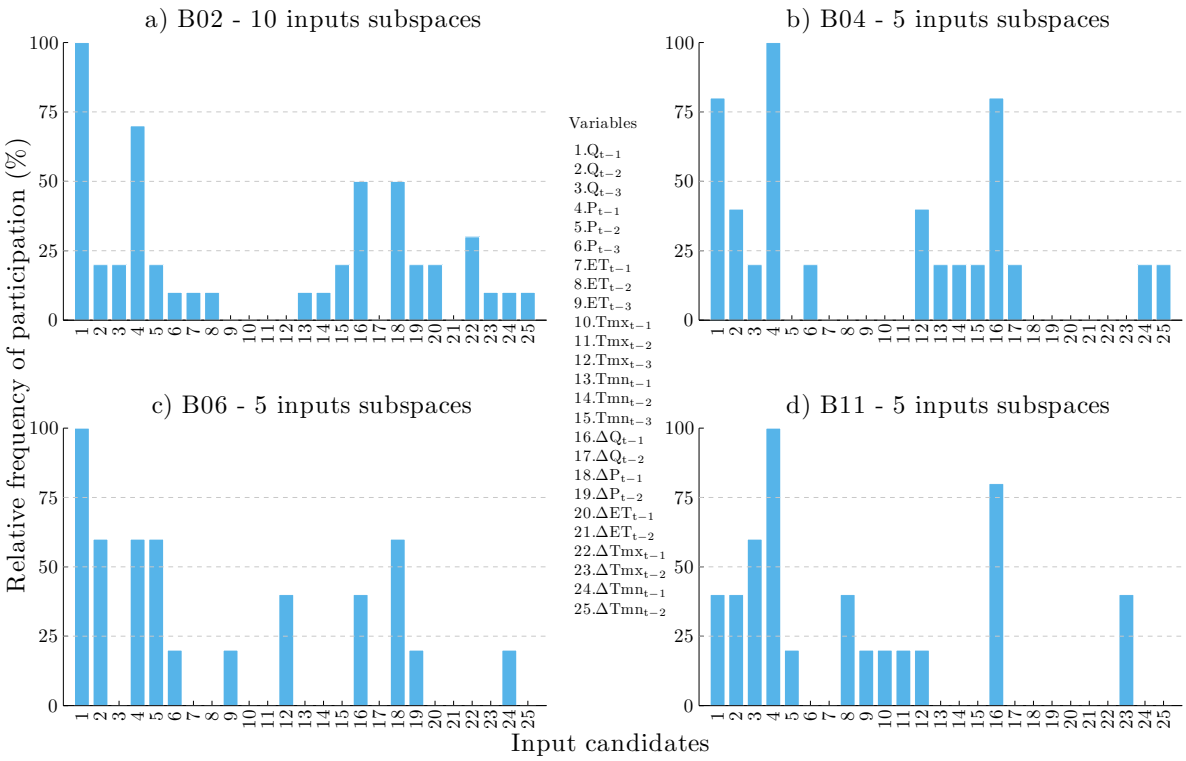


Figure 7.3: Frequency of variable selection found by the DISSENT procedure.

It is important to assess if the initial hypothesis of five inputs in the forecasting model is justified according to the results of the DISSENT methodology. Fig. 7.4 illustrates the median and the interquartile range (iqr) of the 30 evaluations executed in basins B01, B03, B07, B10, and B12. We can see that five variables are sufficient to obtain a stable forecast error, which is the main objective in the search for diversity without degrading the performance of forecasting models. In particular, we can see a greater variability for basins B01 and B12, which is consistent with previous analyzes. Furthermore, the optimal number of input variables is a problem particular to each basin. For example, basins B03 and B07 show a satisfactory and stable performance with only three input variables. However, we use, in all cases, the best five variables, in order to conduct a comparative analysis with respect to the R100P model.

7.3.2 Deterministic evaluation

In order to demonstrate that the low dispersion of the error shown in Fig. 7.4 is not an indicator of low diversity, Fig. 7.5 presents the scatter plot of the prediction ensembles and the observed streamflow for one of the 30 experiments executed on each basin. Except for basins B11 and B12, the mean prediction ensemble is relatively consistent with observations. It can be observed that the prediction ensemble presents high underestimation in the cases of peak events. A detailed evaluation of the hydrographs reveals that such shortcoming is concentrated

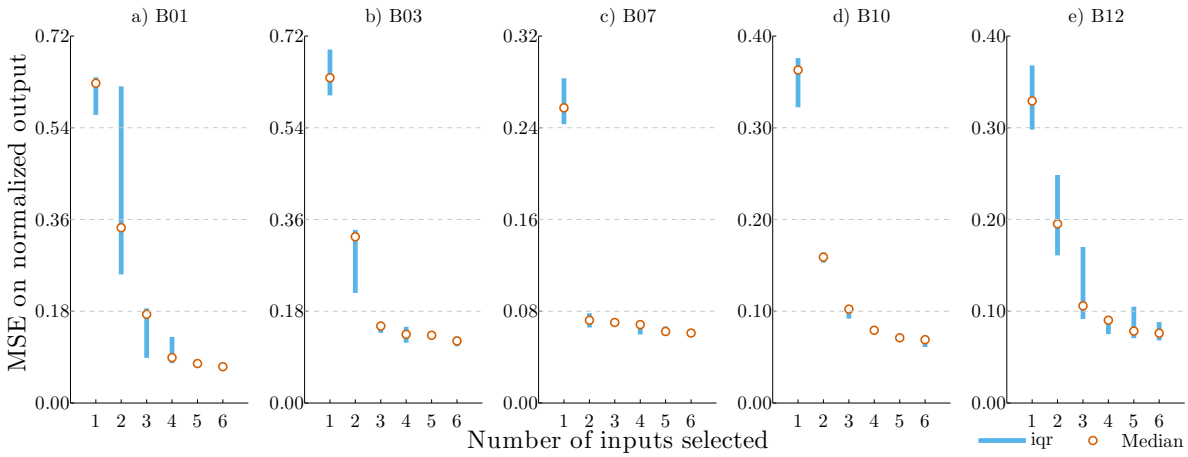


Figure 7.4: Interquartile range (iqr) and median of the ANN ensemble errors according to the different input subspaces.

in peak events prior to a mean or low streamflow. This behaviour is a common problem in most ANN streamflow forecasting models. For example, Abrahart et al. [3] suggested the implementation of an optimization based on a combination of the MSE and a timing correction factor. However, given the achieved ensemble diversity it may be possible to reduce this type of error with a more elaborated voting process than the average value or a post-processing mechanism such as "overproduce and select".

To evaluate the DISSENT efficiency from a deterministic point of view, Table 7.3 presents gains for each basin according to Eq. 6.3. Observe that higher gains are concentrated in basins with the lower NSE and PI criteria, in the same way showed in the stratification analysis that used 100% of data and an optimal stratification scheme (Fig. 6.8). Indeed, the median MSE gain (10.7%) confirms the relevance of the DISSENT methodology in comparison with the optimal stratification scheme using predefined inputs, which shows a MSE gain of −0.22% (Table 7.3).

Basins B11 and B12, identified as more complex to model, are found to largely underestimate high streamflow events, which becomes a critical problem in operational management and decision making activities [3]. One way to address this problem is the evaluation of probabilistic forecasts from a non-reductionist view, consequently, in the next section, we present the probabilistic assessment of the tools described in Sect. 1.3.

7.3.3 Probabilistic evaluation

Concerning the reliability, Fig. 7.6 compares R100P and DISSENT models for basins B02, B05, B07, and B12. We must keep in mind that, in a fully reliable system, the observed conditional probability is equal to the probability in analysis (see Sect. 1.3.3), i.e. all the points in the reliability diagram should fall on the diagonal 1:1 line (dashed line). It is clear

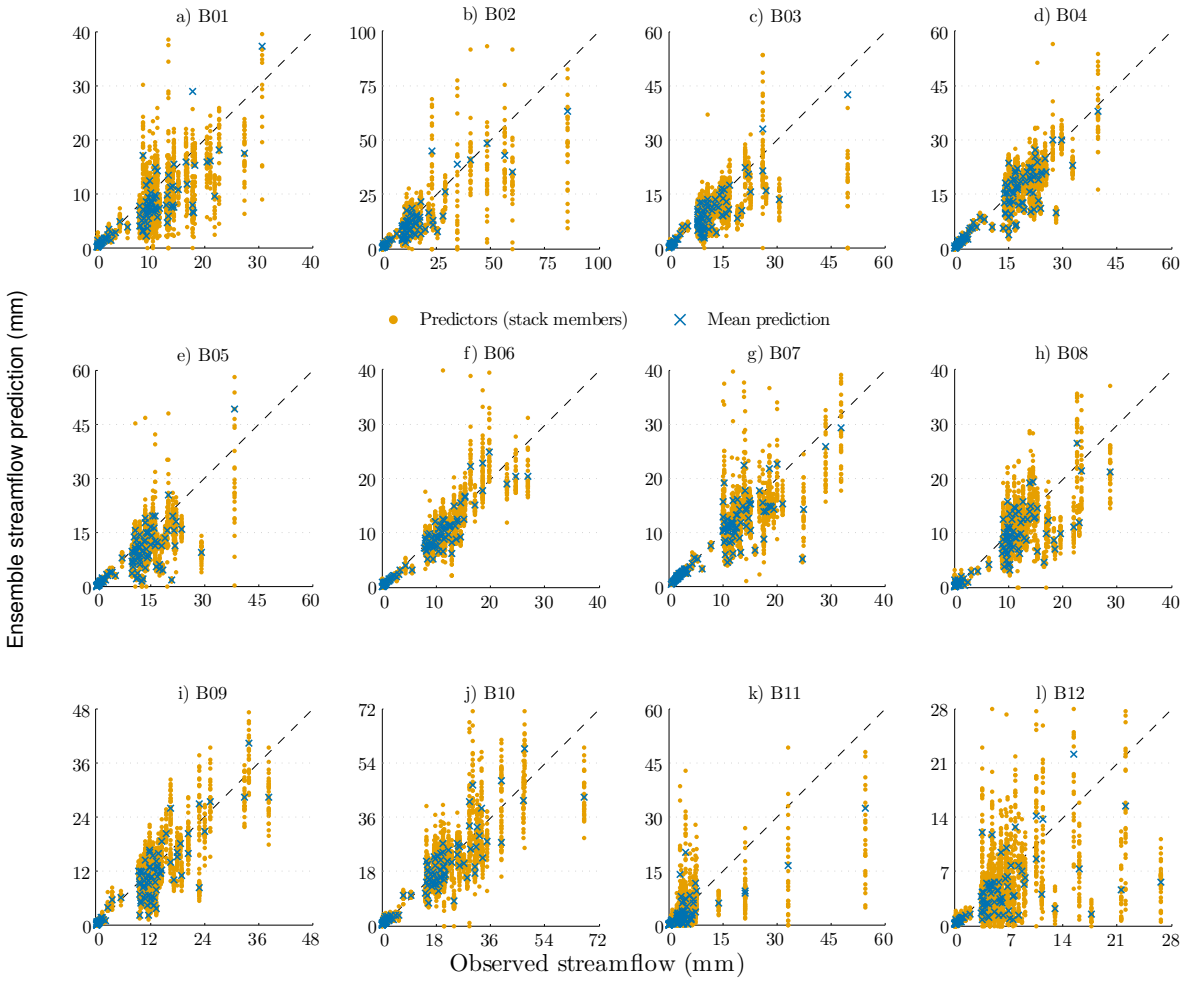


Figure 7.5: Scatter plot of ensemble streamflow prediction and observed streamflow.

in Fig. 7.6 that the R100P model leads to low reliability in basins B02, B05, and B07, since it presents an average forecast that is larger than the average observation (overforecasting), coinciding with the results presented by Boucher et al. [21] in their evaluation of an ANN ensemble using a different database. In contrast, the DISSENT methodology produced more reliable forecasts in basins B02, B05, and, to a lesser extent, in basin B07. The decision to use five input variables instead of three in this basin as suggested by the IVS analysis seems to penalize the performance of the DISSENT methodology.

With regard to basin B12, the reliability diagram shows good calibration with both models, even if in the above sections, we have noted the poorer performance of these models, reflecting the need to complement the probabilistic analysis with other scores. In this sense, the reliability diagram is commonly accompanied by the histogram of ranks or Talagrand diagram (Sect. 1.3.4), which simultaneously reveals characteristics related to the reliability, the bias, and the consistency of the forecast (Fig. 7.7). The goal for rank histograms is to obtain a flat

Table 7.3: Deterministic performance functions to evaluate the DISSENT methodology.

Basin	MAE		MSE		NSE		PI	
	R100P mm	Gain %	R100P mm^2	Gain %	R100P adim.	Gain %	R100P adim.	Gain %
B01	0.301	32.3	0.897	42.0	0.684	19.4	0.286	104.8
B02	0.547	46.2	1.751	47.7	0.707	19.7	0.536	41.3
B03	0.251	9.3	0.731	35.9	0.752	11.8	0.623	21.7
B04	0.443	10.4	0.977	9.2	0.877	1.3	0.705	3.8
B05	0.258	− 2.7	0.780	7.5	0.805	1.9	0.567	5.7
B06	0.148	7.5	0.162	9.0	0.936	0.6	0.715	3.6
B07	0.235	4.6	0.408	5.0	0.882	0.7	0.624	3.0
B08	0.233	5.0	0.506	3.9	0.821	0.9	0.473	4.4
B09	0.288	20.2	0.570	17.0	0.821	3.7	0.563	13.2
B10	0.352	7.9	0.906	12.2	0.898	1.4	0.608	7.9
B11	0.120	− 6.2	0.445	25.6	0.542	21.7	0.500	25.7
B12	0.146	16.1	0.372	7.7	0.544	6.5	0.468	8.9
Median		8.6		10.7		2.8		8.4

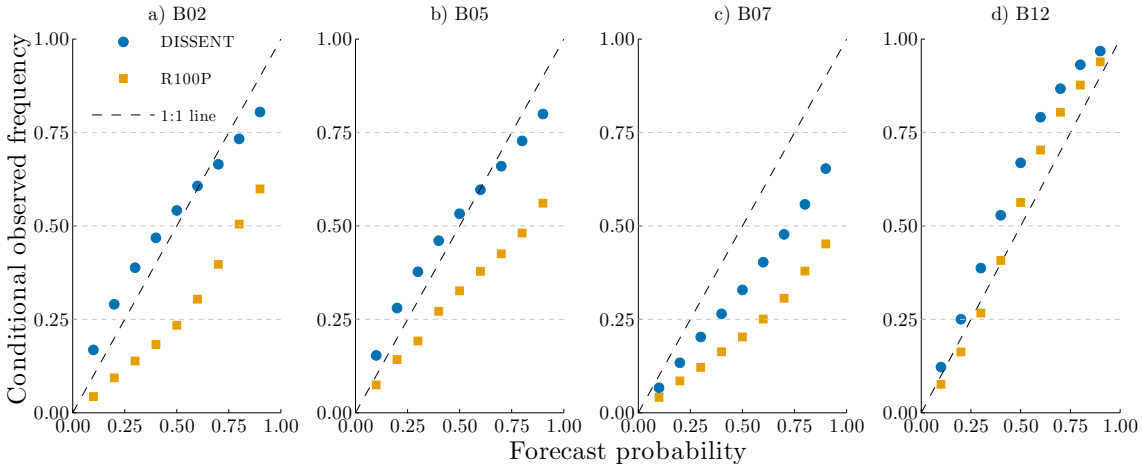


Figure 7.6: Reliability diagrams evaluated in R100P and DISSENT models.

distribution (dashed line). The R100P model results (upper panels) form a U−shaped rank histogram identifying underdispersion (overconfident) in basins B02, B05, and B07, because the ensemble members tend to be similar to each other and different from the observed values (low diversity).

Observed values are too frequently located at the outskirt of the 31 bin ensembles (30 predictors + the observation), so the extreme ranks are overpopulated, and present themselves too rarely as a middle value, so the central ranks are underpopulated. Note that the DISSENT scheme solves this problem partially, as it seeks diversity of prediction. Remark also that the frequency of the first bin in the R100P model is redistributed in the central part of the rank histogram

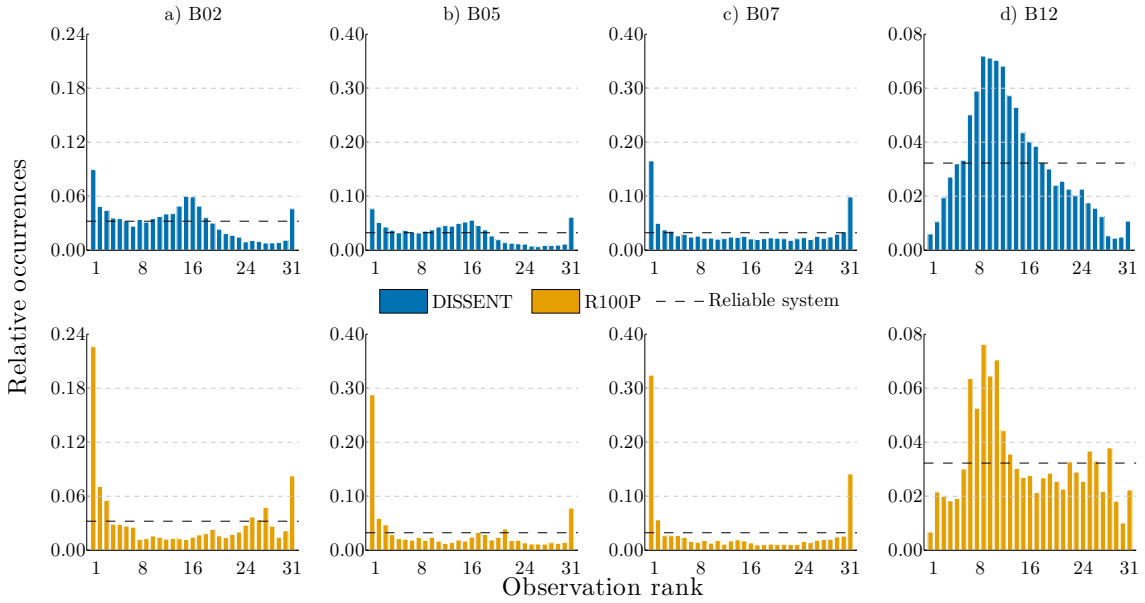


Figure 7.7: Relative rank histograms evaluated in R100P and DISSENT models.

in the DISSENT model. This first bin represents the condition in which all members of the ensemble are greater than the observed value. In spite of such diversity in the ensembles, it still reflects underforecasting in the frequency of the last bin, which represents the case in which the observation is higher than all members of the ensemble.

The proposed multi-criteria evaluation is complemented by the evaluation of the mean CRPS and the mean IGNS. Additionally, comparison of the MAE with the mean CRPS allows us to infer about the relative quality of both the deterministic and the probabilistic systems. So, Table 7.4 presents the MAE as well as the mean CRPS and other scores for the DISSENT evaluations.

Clearly, in all cases, the mean CRPS is lower than the MAE in the DISSENT scheme results, which indicates that the ANN ensemble performs better when taken as a whole than when aggregated in a single averaged predictor. However, the median gain is limited to 8% and the gain is negative (loss) in basin B11.

Regarding the mean IGNS, which heavily penalizes bias ensembles, it shows a median gain around 100% reflecting, to a large extent, the redistribution of the rank histogram illustrated in Fig. 7.7. Despite this, basins B11 and B12 have again poorer results. This is why we adopted the median as a measure of central tendency, accepting that our methodology, based on classical FFNNs, must be complemented with more ANN refinements in the case of semiarid regions, for example considering Echo State Networks (ESNs), such as suggested by Vos [149].

With respect to the MSE evaluated on the reliability diagrams, the median gain reaches 69.5%, which represents a significant improvement in some basins as presented in Fig. 7.6. Regarding

Table 7.4: Probabilistic scores to evaluate the DISSENT methodology relevance.

Basin code	MAE		CRPS		IGNS		RD_{MSE}		δ	
	DISSENT mm	R100P mm	DISSENT mm	Gain %	R100P bits	Gain %	R100P unitless	Gain %	R100P unitless	Gain %
B01	0.204	0.172	0.159	7.5	− 0.698	45.8	1.496	−296.3	51.093	0.3
B02	0.295	0.269	0.238	11.7	4.139	106.8	49.421	90.6	341.975	77.6
B03	0.227	0.189	0.183	3.1	0.952	156.9	15.765	85.5	180.509	72.9
B04	0.397	0.351	0.322	8.4	8.594	80.9	54.493	65.7	241.658	50.3
B05	0.265	0.241	0.216	10.6	3.496	110.1	44.883	90.9	331.388	77.6
B06	0.137	0.118	0.108	8.4	2.105	154.1	26.699	85.6	217.855	68.0
B07	0.224	0.195	0.187	4.4	11.197	82.9	94.241	67.2	682.114	74.3
B08	0.221	0.190	0.178	6.3	0.363	318.0	6.228	− 2.3	133.189	54.1
B09	0.230	0.198	0.178	10.0	− 0.410	96.1	8.834	71.7	248.399	71.4
B10	0.324	0.267	0.240	10.1	3.068	104.0	69.126	91.3	352.189	74.9
B11	0.128	0.073	0.081	−11.8	− 2.308	− 26.4	14.016	− 76.9	394.219	49.9
B12	0.122	0.092	0.091	1.3	− 2.068	− 10.7	3.990	−324.5	115.978	16.3
Median				8.0		100.1		69.5		69.7

the delta ratio, the median gain attains 69.5% reflecting the slight redistribution of bins found for some basins (Fig. 7.7). However, we must consider the higher underdispersion of the reference model (R100P), so more efforts should be considered to achieve better results on this score, which has a strong influence on the other properties of probabilistic prediction, as we discussed in precedent chapters.

7.4 Conclusion and future work

The active pursuit of diversity in the construction of a HEPS without degrading the performance of each member of the ensemble is a challenge since each member must have a similar reliability in order to achieve a HEPS with high consistency. In our case, the HEPS was primarily devised following three guidelines:

- Exploiting the duality between "instability" and "precision" of the ANN;
- Promoting and verifying that each member of the ensemble has high performance; and
- Identifying patterns contained in the information to force the complementarity of the members.

These guidelines largely coincide with the principles of the AdaBoost model [63, 136], which is based on the combination of weak models that recurrently are specialized in certain types of domain regions.

In the context of hydrological forecasts, as well as in machine learning, the diversity of ensembles is usually explored in domains where the uncertainty is evident. These domains take into account the input information (data level), the related variables (input space level), the

model conceptualization (model level), the parameters of such models (parametric level), and finally the decision processes that justify the selection or fusion of the predictors (model combiner level). In this study, we investigated the propagation of three uncertainty sources: data selection, input space, and parametric levels. More specifically, we analyzed the influence of pattern selection (stratification) and its impact on the formulation of a dynamic input space.

We then evaluated a procedure called DISSENT, in which each ANN member or individual predictor was trained based on input spaces defined by the application of a stepwise IVS on different stratified sub-samples. These results confirmed an appreciable gain while only considering the data variability. However, the main advantages of this framework occurred in the probabilistic prediction model, which unlike the classic reductionism of the deterministic approach, strongly penalizes the lack of diversity in ensembles. In general, the baseline model presented a low reliability, in agreement with Boucher et al. [21], who, in a probabilistic ANN context, evaluated the possibility of improving the diversity of optimized models in each training epoch of the ANN. In our case, reliability was greatly improved with a gain of 70%. Regarding the CRPS and IGNS, results showed a gain of 8% and 100%, respectively. Furthermore, DISSENT also improved the strong underdispersion of the R100P model, as depicted by the rank histogram.

This work prompted the following thoughts for future works:

- Establish a basic hydrological criteria that may be associated with the level of complexity required for ANN models. We showed that the aridity index and the coefficient of variation of maximum annual streamflow series can be associated to poor performance of FFNNs, which are adequate as prediction model in most of the cases.
- The CRPS is commonly used in meteorology as a multipurpose score, simultaneously evaluating bias, reliability, and resolution. Results produced here showed that the CRPS does not respond much to diversity. It would therefore be interesting to better assess, theoretically and experimentally, the relationship of the CRPS with diversity in prediction ensembles . It is easy to note that the CRPS reaches its minimum value when all predictors of an ensemble are equal and coincide with the observation; a situation that is hypothetically impossible to accomplish.
- Given the length of the training dataset (18 years), it is necessary to perform sensitivity analyzes to establish the generality of the proposed methodology regarding this aspect.
- Given the large availability of ANN models [97, 98, 149], it would be interesting to consider the inclusion of an ANN ensemble with members of different structures (models level) to assess the level of diversity associated with the structure of the prediction model, as suggested by Abrahart et al. [6] and Brochero et al. [30].
- It could be interesting to better assess the relevance of selection or fusion methods, such as those discussed by Kuncheva [92] – the combiner level. The "overproduce and select" philosophy can easily be adapted in this context, as proposed in precedent chapters.

Part IV

Conclusion, Contributions, and Future Work

Conclusion

The relationships between diversity and the performance of Hydrological Ensemble Prediction Systems (HEPSs) motivated this research in two meaningful directions:

1. Optimization process based on the selection of the 'best' predictors of a predefined 800-member HEPS. The latter was configured using 16 lumped hydrological models driven by the 50-member weather ensemble forecasts from the European Centre for Medium-range Weather Forecasts (ECMWF) Ensemble Prediction System (EPS); and
2. Diversity in an Artificial Neural Network (ANN) ensemble defined from uncertainty conceptualization at different levels.

For the HEPS optimization, diversity was explored through the selection of a subset of predictors that represent "sufficient" system diversity (Chap. 3, 4, and 5). It is important to note the analogy between the Input Variable Selection (IVS) problem and the selection task evaluated here, because instead of picking the most important model inputs, we seek predictors that optimize the probabilistic HEPS output. For this, we use Backward Greedy Selection (BGS) combined with Cross-Validation (CV) (BGS-CV) and some probabilistic scores.

However, this problem required to hypothesize (later validated from results) that each hydrological member is a variable for subsequent interpretation in terms of Hydrological Models Participation (HMP). It is very important to highlight that member selection was not performed on equiprobable meteorological members but rather on the 800-member hydrological response.

In Chap. 3, we were able to observe the pronounced interaction between the IGNorance Score (IGNS), the error on the reliability diagram, the δ ratio, and, to a lesser extent, the Continuous Ranked Probability Score (CRPS). Therefore, the design of a Combined Criterion (CC) led to an important methodological improvement that integrated many characteristics of each score. The δ ratio turned out as the best single optimization criterion, close to the CC.

We explored the BGS-CV with the CC in Chap. 4, demonstrating the generalization ability of this scheme for other forecast time horizons and neighbouring basins. A regional integration scheme was proposed. Moreover, the best balance of scores was achieved with a number of members fluctuating between 30 and 100, maximizing the quality of the system in terms of reliability, consistency, and resolution.

Chapter 5 was the object of a general framework in which the selection of members, expressed as the HMP, directly oriented the evaluation of representative precipitation members at each time step to subsequently propagate them into their respective hydrological model. Note that in Chap. 3 and 4, the HMP was tested with randomly picked meteorological members. Additionally, we explored four selection techniques to obtain a 48-member HEPS for the sake of a fair comparison with a reference model called the uniform HMP scheme that evaluated

the propagation of three representative meteorological members into 16 hydrological models. Three of the selected techniques, usually employed in an IVS context, were Linear Correlation Elimination (LCE), Mutual Information (MI), and BGS. The fourth one, commonly used in a multi-objective optimization, was the Nondominated Sorting Genetic Algorithm II (NSGA-II). Results showed that difficulties in simplifying mainly originated from the preservation of the system reliability. Compared with the efficiency shown by BGS and NSGA-II, both the uniform HMP scheme and the simplification schemes based on members' correlation (LCE, MI) showed generally poor performance. We highlighted the advantages of using NSGA-II because it allowed a direct trade-off among the evaluated scores (also containing BGS selections) and it was about 5 times faster than BGS.

We consider important to stress that the methods evaluated in Chap. 5 are fully transferable to the model response level evaluated in ANN ensembles (Chap. 6 and 7).

In order to analyze the HEPS diversity in an ANN ensemble (Chap. 6 and 7), we explored three uncertainty levels: datasets, input sub-spaces, and the inherent variability of the ANN gradient-based training process. In this case, we used 12 basins originating from the second and third workshops of the MOdel Parameter Estimation eXperiment (MOPEX) project [53]. Deterministic and probabilistic results revealed the higher performance of the proposed methodology when compared to an Ensemble of 30 FFNNs trained with early stopping using a **R**andom sampling of **100 P**ercent of the available information and a single predefined set of inputs variables (R100P).

The R100P model performance confirmed the general good performance of ANNs, except for one basin that turned out to be atypical, and two more that embodied the difficulty of streamflow prediction in semiarid areas. This latter condition did not improve substantially with the proposed methodology, coinciding with the difficulties found by Vos [149] in the application of complex ANN structures using the same database. The performances of the stratification resampling scenarios ended up similar to the R100P one, but using only half of the database. The main advantage of this procedure consisted in the efficient selection of the training information with stratified sub-samples. We thus promoted ANN ensembles in which each predictor is trained based on input spaces defined by an IVS application on different stratified sub-samples. This novel method, called Dynamic Input Spaces imposed by Stratified Examples propagated on artificial Neural networks Training (DISSENT), led to a gain of 8%, 100.1%, 69.5%, and 69.7%, with respect to the mean CRPS, the mean IGNS, the Mean Square Error (MSE) evaluated in the Reliability Diagram (RD_{MSE}), and the δ ratio, respectively.

Overall, we conclude that ensemble diversity must be seen as complementarity between predictors, which requires proper tuning in terms of the objectives of the probabilistic prediction. In such context, machine learning would usually focus on combining predictors in order to op-

timize some deterministic criterion like bias, considered as one of the main features to preserve in probabilistic forecasting. Accordingly, "sufficient" diversity can be obtained from the active integration of scores in the ANN training itself or in a multi-score framework as presented above.

Contribution

Even if detailed findings were identified at the end of each chapter, we will use a few more pages to highlight the contributions originating from this thesis and propose some guidelines for future work. They are grouped as conceptual background, knowledge transfer between the machine learning and hydrometeorology communities, ANN applications, and machine learning applications for HEPS post-processing.

Contribution to conceptual background

- Development of the concept of diversity for Hydrological Ensemble Prediction Systems: in the machine learning community, the advancement of this concept, from the evolution of the bias-variance dilemma, the accuracy-diversity breakdown, and the bias-variance-covariance decomposition, led naturally to the evaluation of multiple scenarios. This is a contribution that is actively sought after for probabilistic prediction in the hydrometeorological community.
- Identification of model response combiners: the parallelization of schemes of uncertainty observed in physical conceptualization and in mathematical modelling allowed to emphasize the importance of modelling a combined response accounting for the uncertainty associated with meteorological variables, uncertainty in the conceptualization of the hydrologic and hydraulic processes, and parametric uncertainty in the hydraulic and hydrologic models (see Cascading model uncertainty presented by Pappenberger et al. [115]). From a mathematical point of view, uncertainty originates from the data, the input subsets, the models, the parameter settings, and the final response model combiners.
- Implementation of a novel framework in a HEPS simplification process: in this framework, the simplification process is conceived without sacrificing forecast quality through the identification of predictors that bring the greatest contribution to a simplified HEPS.
- Implementation of a basic criterion called the MeDian of the Coefficients of Variation (MDCV) to infer the relationship between the dispersion and other characteristics of probabilistic prediction.
- Generation of a combined criterion that merges the normalized results of the CRPS, the IGNS, the Reliability Diagram (RD), the rank histogram (δ ratio), and the MDCV: the importance of this approach lies in its possible inclusion in tools of mono-objective optimization. Also, in many situations, improving one individual score is achieved at the expense of another one.

- Design of a methodology for a nonparametric evaluation of the optimal Hydrological Models Participation (HMP) as an indicator of the number of representative meteorological members to propagate into each hydrological model.
- Evaluation of the shortcomings of an intuitive simplification scheme based on members correlation.
- Design of a multi-score framework that identifies trade-offs between them, facilitating decision making scenarios according to the properties that could be prioritized in particular cases.

Contribution to knowledge transfer

- Extension of probabilistic scores developed by the hydrometeorological community to the machine learning community: although the latter already had experienced with several functions associated to the diversity and system entropy concepts, usage of the CRPS, the IGNS, the reliability diagram, and the delta ratio is still uncommon for them.
- Inclusion in the proposed methodologies of the generalization ability concept, which is a major ambition of machine learning: this concept is defined as the capacity to simulate output from examples that differ from those used in training, relating the quantity and quality of the information, and the bias and model variance.
- Implementation of procedures to prevent overfitting: in fact, overfitting is possible even when the calibration data are noise-free, especially when a relatively small number of examples are used for calibrating a relatively large number of model parameters. To avoid overfitting, machine learning has developed techniques such as regularization and early stopping [7, 74]. This problem should not be seen as insignificant in the calibration of conventional hydrological models.
- Introduction of CV (typical in the machine learning community) to evaluate the complexity needed in a hydrological model: we exploited it with BGS. Although this method is very intuitive since it allows training and testing with different datasets, it is generally not included in the standard calibration of hydrological models.
- Proposal of the DISSENT framework for ANN and the multi-model 800-member HEPS: it favours implicit diversity in ensemble performance both in the machine learning and hydrometeorological communities.

Contributions to ANN applications

- Active inclusion of implicit diversity in the configuration of an ANN ensemble, without sacrificing the accuracy of individual members.
- Generation of an experimental protocol that allows simple verification of results: it begins with a clear definition of the datasets used, data preparation, network configuration, and

training and stopping criteria. It also opened the possibility of comparative framework results using open access datasets.
- Implementation of a hydrological criterion, such as the aridity index and the proposed coefficient of variation of annual maximum streamflows, to justify the implementation of complex ANN structures.
- Evaluation of multiple scenarios of stratification that demonstrate that the informativeness is far more important than the volume of data dedicated to ANN training.
- Implementation of a multi-criteria framework, in which we evaluate deterministic criteria as multiple probabilistic scores: in this regard, there are few studies that include the Persistence Index (PI) as function error. It is particularly well designed for prediction evaluation considering that the last observed streamflow is generally one of the ANN inputs. Similarly, the probabilistic evaluation of the ANN ensembles with scores developed by the hydrometeorological community is in its infancy.
- Development of the DISSENT methodology based on the relevance and simplicity of three techniques: k-means clustering, Forward Greedy Selection (FGS), and Feed-Forward Neural Network (FFNN).

Contributions to HEPS post-processing

- Implementation of a member selection framework that uses a variation of the k-fold CV, a multi-score approach, and a selection method called BGS: additionally, this framework allows inference about the optimal number of members to be selected based on a simple relationship between the ensemble size and its performance.
- Adaptation of clustering tools in the proposed mechanism for the integration of members selection.
- Adaptation of a filter tool for MI in the context of CC minimization: we propose a linear search for the best combination between the parametrization proposed to evaluate the MI and the number of quantile defined in the discretization step.
- Adaptation of the NSGA-II technique in the context of selecting members.

Future work

It is still important to decipher the relationship between diversity and the hydrological model complementarity. In a next phase, an explanation should be obtained, from a hydrological point of view, for the processes that lead to diversity in terms of catchment structure. This task is more difficult in ANN, considering their black-box nature, than for the structural evaluation of hydrological models used in the the 800-member HEPS. However, the water resources ANN community is already aware that one of the paradigms of operational systems is to establish relationships between the variables, in a cause-effect liaison that guarantees a certain level of security in the decision making stage.

It is precisely at this operational level and decision making scenario that we must set objectives for including hydroinformatics tools for the optimization of the quality of predictions. The relevance and value of such processes become apparent as soon as it becomes possible to quantify its impact on real applications.

With respect to the selection of members in the complex 800-member HEPS, we suggest prioritizing the enrichment of the databases by including a larger number of events, which would allow inferring about the behaviour of member selection based on event types. In this regard, it is clear that the evaluation of the larger events becomes more important in an operational context.

More specifically, we consider that in probabilistic prediction with ANN ensembles the following guidelines should be further evaluated:

- Inclusion of probabilistic scores in the ANN training;
- Inclusion of different ANN structures to consider uncertainty due to the inductive bias of the model; and
- Evaluation of member selection methodologies in a "overproduce and select" framework.

Appendix A

Publications Resulting from this Thesis

The work resulting from these investigations has been published in journal articles and conference proceedings.

Journal articles

- D. Brochero, F. Anctil, and C. Gagné (2011b). "Simplifying a hydrological ensemble prediction system with a backward greedy selection of members – Part 1: Optimization criteria". In: *Hydrol. Earth Syst. Sci.* 15.11, pp. 3307–3325.
- D. Brochero, F. Anctil, and C. Gagné (2011a). "Simplifying a hydrological ensemble prediction system with a backward greedy selection of members – Part 2: Generalization in time and space". In: *Hydrol. Earth Syst. Sci.* 15.11, pp. 3327–3341.

Papers in conference proceedings

- F. Anctil, D. Brochero, and C. Gagné (2011). *Which Optimization Criterion Leads to the Reliable Simplification of a Hydrological Ensemble Prediction System with a Backward Greedy Selection of Members?* In: *European Geosciences Union (EGU), Geophysical Research Abstract.* Vol. 13. Vienna, Austria.
- D. Brochero, F. Anctil, and C. Gagné (2011c). *An experience on the selection of members for simplifying a multimodel hydrological ensemble prediction system*. In: *CSHS Workshop: Operational River Flow and Water Supply Forecasting.* Vancouver, Canada.
- D. Brochero, F. Anctil, and C. Gagné (2012b). *Forward Greedy ANN input selection in a stacked framework with Adaboost.RT – A streamflow forecasting case study exploiting radar*

rainfall estimates. In: *European Geosciences Union (EGU), Geophysical Research Abstract*. Vol. 14. Vienna, Austria.

- D. Brochero, F. Anctil, and C. Gagné (2012a). *Comparison of three methods for the optimal allocation of hydrological model participation in an Ensemble Prediction System*. In: *European Geosciences Union (EGU), Geophysical Research Abstract*. Vol. 14. Vienna, Austria.
- D. Brochero, F. Anctil, K. López, and C. Gagné (2013c). *Finding diversity for building one-day ahead Hydrological Ensemble Prediction System based on artificial neural network stacks*. In: *European Geosciences Union (EGU), Geophysical Research Abstract*. Vol. 15. Vienna, Austria.
- D. Brochero, F. Anctil, and C. Gagné (2013b). *Evolutionary Multiobjective Optimization for Selecting Members of an Ensemble Streamflow Forecasting Model*. In: *Proceedings of the Genetic and Evolutionary Computation Conference*. Amsterdam, Netherlands.

Journal article - in preparation

- D. Brochero, K. López, F. Anctil, and C. Gagné (2013e). *Stratification analysis in Artificial Neural Networks for streamflow forecasting*. In preparation.
- D. Brochero, F. Anctil, and C. Gagné (2013a). *Diversity in Artificial Neural Networks ensembles with applications to streamflow predictions*. In preparation.
- D. Brochero, F. Anctil, and C. Gagné (2013d). *Hydrological models weight evaluation and representative meteorological members propagation to orient a multimodel Hydrological Ensemble Prediction System optimization*. In preparation.

Appendix B

MSE Decomposition - Deterministic Case

Formally, the decomposition can be proved adding and subtracting the mean forecast and the mean observations ($\bar{\mathbf{y}} = \sum_{t=1}^{N} y^t, \bar{\mathbf{o}} = \sum_{t=1}^{N} o^t$, respectively), and completing the squaring process within the brackets given that $(\bar{\mathbf{y}} - \bar{\mathbf{o}})$ is constant, $\sum\sum(o_i - \bar{\mathbf{o}}) = \sum\sum(y_i - \bar{\mathbf{y}}) = 0$.

$$
\begin{aligned}
\mathrm{MSE}(\mathbf{o}, \mathbf{y}) &= \frac{1}{n}\sum_{i=1}^{N} (y_i - o_i)^2 \\
&= \frac{1}{n}\sum_{i=1}^{N} \left[(y_i - \bar{\mathbf{y}}) - (o_i - \bar{\mathbf{o}}) + (\bar{\mathbf{y}} - \bar{\mathbf{o}})\right]^2 \\
&= \frac{1}{n}\sum_{i=1}^{N} \left[(y_i - \bar{\mathbf{y}}) - (o_i - \bar{\mathbf{o}})\right]^2 + (\bar{\mathbf{y}} - \bar{\mathbf{o}})^2 + 2\left[(y_i - \bar{\mathbf{y}}) - (o_i - \bar{\mathbf{o}})\right](\bar{\mathbf{y}} - \bar{\mathbf{o}}) \\
&= \frac{1}{n}\sum_{i=1}^{N} (y_i - \bar{\mathbf{y}})^2 + (o_i - \bar{\mathbf{o}})^2 - 2(y_i - \bar{\mathbf{y}})(o_i - \bar{\mathbf{o}}) + (\bar{\mathbf{y}} - \bar{\mathbf{o}})^2 \\
\mathrm{MSE}(\mathbf{o}, \mathbf{y}) &= \underbrace{\frac{1}{n}\sum_{i=1}^{N} (y_i - \bar{\mathbf{y}})^2}_{\text{forecasts variance}} + \underbrace{\frac{1}{n}\sum_{i=1}^{N} (o_i - \bar{\mathbf{o}})^2}_{\text{observations variance}} + \underbrace{(\bar{\mathbf{y}} - \bar{\mathbf{o}})^2}_{\text{bias}} \\
&\quad - \underbrace{2\frac{1}{n}\sum_{i=1}^{N} (y_i - \bar{\mathbf{y}})(o_i - \bar{\mathbf{o}})}_{\text{covariance}}
\end{aligned}
\tag{B.1}
$$

Appendix C

MSE Decomposition - Expected Square Error

Adding and subtracting $E[y(x)]$, and completing the squaring processing:

$$\begin{aligned}
E_\chi\left[\left(E[o|x]-y(x)\right)^2|x\right] &= E_\chi\left[\left((E[o|x]-E[y(x)])-(y(x)-E[y(x)])\right)^2|x\right] \\
&= E_\chi\Big[\Big((E[o|x]-E[y(x)])^2+(y(x)-E[y(x)])^2 \\
&\quad -2\left(E[o|x]-E[y(x)]\right)\left(y(x)-E[y(x)]\right)\Big)|x\Big] \\
&= \underbrace{E_\chi\left[\left(E[o|x]-E[y(x)]\right)^2|x\right]}_{\text{bias}^2}+\underbrace{E_\chi\left[\left(y(x)-E[y(x)]\right)^2|x\right]}_{\text{variance}} \\
&\quad -2\underbrace{\left(E[o|x]-E[y(x)]\right)}_{\text{constant}}\underbrace{E_\chi\left(y(x)-E[y(x)]\right)|x}_{=0}
\end{aligned} \tag{C.1}$$

Appendix D

MSE Decomposition - Multimodel Approach

Brown [36] presented the following proof of the ambiguity decomposition at a single datapoint (y_d is short for $y_d(\mathbf{x}^t)$, $\bar{y}$ for $\bar{y}(\mathbf{x}^t)$, and o for o^t):

$$\begin{aligned}
\sum_{d=1}^{D}(y_d-o)^2 &= \sum_{d=1}^{D}(y_d-\bar{y}+\bar{y}-o) \\
&= \sum_{d=1}^{D}\Big((y_d-\bar{y})^2+(\bar{y}-o)^2-2\,(y_d-\bar{y})\,(\bar{y}-o)\Big) \\
&= \sum_{d=1}^{D}(y_d-\bar{y})^2+\sum_{d=1}^{D}(\bar{y}-o)^2-2\,(\bar{y}-o)\underbrace{\sum_{d=1}^{D}(y_d-\bar{y})}_{=0} \\
\sum_{d=1}^{D}(\bar{y}-o)^2 &= \sum_{d=1}^{D}(y_d-o)^2-\sum_{d=1}^{D}(y_d-\bar{y})^2
\end{aligned} \tag{D.1}$$

Bibliography

[1] Abrahart, R., See, L., and Dawson, C. (2008a). "Neural Network Hydroinformatics: Maintaining Scientific Rigour". In: *Practical Hydroinformatics*. Vol. 68. Springer Berlin Heidelberg, pp. 33–47 (cit. on pp. 14, 107).

[2] Abrahart, R. J. and See, L. (2000). "Comparing neural network and autoregressive moving average techniques for the provision of continuous river flow forecasts in two contrasting catchments". In: *Hydrol. Processes* 14.11-12, pp. 2157–2172 (cit. on pp. 107, 124, 129).

[3] Abrahart, R. J., Heppenstall, A. J., and See, L. M. (2007). "Timing error correction procedure applied to neural network rainfall—runoff modelling". In: *Hydrol. Sci. J.* 52.3, pp. 414–431 (cit. on p. 135).

[4] Abrahart, R. J., See, L., and Solomatine, D. (2008b). *Practical Hydroinformatics. Computational Intelligence and Technological Developments in Water Applications*. Vol. 68. Springer Berlin Heidelberg (cit. on p. 2).

[5] Abrahart, R. J., See, L. M., Dawson, C. W., Shamseldin, A. Y., and Wilby, R. L. (2010). *Nearly two decades of neural network hydrologic modeling*. In: *Advances in Data-Based Approaches for Hydrologic Modeling and Forecasting*. World Scientific Publishing. Chap. 6, pp. 267–346 (cit. on p. 9).

[6] Abrahart, R. J., Anctil, F., Coulibaly, P., Dawson, C. W., Mount, N. J., See, L. M., Shamseldin, A. Y., Solomatine, D. P., Toth, E., and Wilby, R. L. (2012). "Two decades of anarchy? Emerging themes and outstanding challenges for neural network river forecasting". In: *Prog. in Phys. Geogr.* 36.4, pp. 480–513 (cit. on pp. 9, 14, 128, 140).

[7] Alpaydin, E. (2010). *Introduction to Machine Learning*. 2nd ed. The MIT Press (cit. on pp. 5, 21, 32–34, 38, 39, 64, 113, 126, 146).

[8] Anctil, F. and Lauzon, N. (2004). "Generalisation for neural networks through data sampling and training procedures, with applications to streamflow predictions". In: *Hydrol. Earth Syst. Sci.* 8.5, pp. 940–958 (cit. on pp. 10, 11, 21, 106, 107, 124, 126, 129).

[9] Anctil, F. and Rat, A. (2005). "Evaluation of Neural Network Streamflow Forecasting on 47 Watersheds". In: *J. Hydrol. Eng.* 10.1, pp. 85–88 (cit. on p. 127).

[10] Anctil, F., Lauzon, N., and Filion, M. (2008). "Added gains of soil moisture content observations for streamflow predictions using neural networks". In: *J. Hydrol.* 359, pp. 225–234 (cit. on pp. 10, 115, 127).

[11] Anctil, F., Perrin, C., and Andréassian, V. (2004). "Impact of the length of observed records on the performance of ANN and of conceptual parsimonious rainfall-runoff forecasting models". In: *Environ. Modell. Softw.* 19.4, pp. 357–368 (cit. on pp. 37, 107, 113–115).

[12] Anctil, F., Lauzon, N., Andréassian, V., Oudin, L., and Perrin, C. (2006). "Improvement of rainfall-runoff forecasts through mean areal rainfall optimization". In: *J. Hydrol.* 328.3–4, pp. 717–725 (cit. on p. 40).

[13] Anctil, F., Filion, M., and Tournebize, J. (2009). "A neural network experiment on the simulation of daily nitrate-nitrogen and suspended sediment fluxes from a small agricultural catchment". In: *Ecol. Model.* 220.6, pp. 879–887 (cit. on pp. 127, 131).

[14] Anctil, F., Brochero, D., and Gagné, C. (2011). *Which Optimization Criterion Leads to the Reliable Simplification of a Hydrological Ensemble Prediction System with a Backward Greedy Selection of Members?* In: *European Geosciences Union (EGU), Geophysical Research Abstract.* Vol. 13 (cit. on p. 149).

[15] Anderson, J. L. (1996). "A method for producing and evaluating probabilistic forecasts from ensemble model integrations". In: *J. Climate* 9.7, pp. 1518–1530 (cit. on pp. 29, 56).

[16] Arora, V. K. (2002). "The use of the aridity index to assess climate change effect on annual runoff". In: *J. Hydrol.* 265.164, pp. 164 –177 (cit. on p. 116).

[17] Bao, H.-J., Zhao, L.-N., He, Y., Li, Z.-J., Wetterhall, F., Cloke, H. L., Pappenberger, F., and Manful, D. (2011). "Coupling ensemble weather predictions based on TIGGE database with Grid-Xinanjiang model for flood forecast". In: *Adv. Geosci.* 29, pp. 61–67 (cit. on p. 9).

[18] Battiti, R. (1994). "Using mutual information for selecting features in supervised neural net learning". In: *IEEE Trans. on Neural Netw.* 5.4, pp. 537–550 (cit. on p. 89).

[19] Beven, K. and Binley, A. (1992). "The future of distributed models: Model calibration and uncertainty prediction". In: *Hydrol. Processes* 6, pp. 279–298 (cit. on pp. 9, 32, 46).

[20] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., p. 738 (cit. on pp. 32, 35, 36).

[21] Boucher, M.-A., Perreault, L., and Anctil, F. (2009). "Tools for the assessment of hydrological ensemble forecasts obtained by neural networks". In: *J. Hydroinf.* 11.3-4, pp. 297–307 (cit. on pp. 14, 59, 76, 125, 126, 136, 140).

[22] Boucher, M.-A., Laliberté, J.-P., and Anctil, F. (2010). "An experiment on the evolution of an ensemble of neural networks for streamflow forecasting". In: *Hydrol. Earth Syst. Sci.* 14.3, pp. 603–612 (cit. on pp. 11, 14, 27).

[23] Bougeault, P., Toth, Z., Bishop, C., Brown, B., Burridge, D., Chen, D. H., Ebert, B., Fuentes, M., Hamill, T. M., Mylne, K., Nicolau, J., Paccagnella, T., Park, Y.-Y., Parsons, D., Raoult, B., Schuster, D., Dias, P. S., Swinbank, R., Takeuchi, Y., Tennant, W., Wilson, L., and Worley, S. (2010). "The THORPEX Interactive Grand Global Ensemble". In: *Bull. Am. Meteorol. Soc.* 91.8, pp. 1059–1072 (cit. on pp. 10, 46).

[24] Bowden, G. J., Dandy, G. C., and Maier, H. R. (2005). "Input determination for neural network models in water resources applications. Part 1 – background and methodology". In: *J. Hydrol.* 301, pp. 75–92 (cit. on p. 127).

[25] Bowden, G. J., Maier, H. R., and Dandy, G. C. (2012). "Real-time deployment of artificial neural network forecasting models: Understanding the range of applicability". In: *Water Resour. Res.* 48.10, n/a–n/a (cit. on p. 106).

[26] Brochero, D., Anctil, F., and Gagné, C. (2011a). "Simplifying a hydrological ensemble prediction system with a backward greedy selection of members – Part 2: Generalization in time and space". In: *Hydrol. Earth Syst. Sci.* 15.11, pp. 3327–3341 (cit. on p. 149).

[27] — (2011b). "Simplifying a hydrological ensemble prediction system with a backward greedy selection of members – Part 1: Optimization criteria". In: *Hydrol. Earth Syst. Sci.* 15.11, pp. 3307–3325 (cit. on pp. 126, 149).

[28] Brochero, D., Anctil, F., and Gagné, C. (2011c). *An experience on the selection of members for simplifying a multimodel hydrological ensemble prediction system*. In: *CSHS Workshop: Operational River Flow and Water Supply Forecasting* (cit. on pp. 126, 149).

[29] — (2012a). *Comparison of three methods for the optimal allocation of hydrological model participation in an Ensemble Prediction System*. In: *European Geosciences Union (EGU), Geophysical Research Abstract*. Vol. 14 (cit. on pp. 126, 150).

[30] — (2012b). *Forward Greedy ANN input selection in a stacked framework with Adaboost.RT – A streamflow forecasting case study exploiting radar rainfall estimates*. In: *European Geosciences Union (EGU), Geophysical Research Abstract*. Vol. 14 (cit. on pp. 127, 140, 149).

[31] — (2013a). *Diversity in Artificial Neural Networks ensembles with applications to streamflow predictions*. In preparation (cit. on p. 150).

[32] — (2013b). *Evolutionary Multiobjective Optimization for Selecting Members of an Ensemble Streamflow Forecasting Model*. In: *Proceedings of the Genetic and Evolutionary Computation Conference* (cit. on p. 150).

[33] Brochero, D., Anctil, F., López, K., and Gagné, C. (2013c). *Finding diversity for building one-day ahead Hydrological Ensemble Prediction System based on artificial neural network stacks*. In: *European Geosciences Union (EGU), Geophysical Research Abstract*. Vol. 15 (cit. on p. 150).

[34] Brochero, D., Anctil, F., and Gagné, C. (2013d). *Hydrological models weight evaluation and representative meteorological members propagation to orient a multimodel Hydrological Ensemble Prediction System optimization*. In preparation (cit. on p. 150).

[35] Brochero, D., López, K., Anctil, F., and Gagné, C. (2013e). *Stratification analysis in Artificial Neural Networks for streamflow forecasting*. In preparation (cit. on p. 150).

[36] Brown, G. (2004). *Diversity in Neural Network Ensembles*. PhD thesis. School of Computer Science, University of Birmingham (cit. on pp. 4, 21, 22, 126, 155).

[37] — (2009). *A New Perspective for Information Theoretic Feature Selection*. In: *12th International Conference on Artificial Intelligence and Statistics*. Vol. 5, pp. 49–56 (cit. on pp. 38, 39, 87–89, 127).

[38] Brown, G. (2010). "Ensemble Learning". In: *Encyclopedia of Machine Learning*. Springer US, pp. 312–320 (cit. on pp. 22, 23).

[39] Buizza, R. (2005). "EPS skill improvements between 1994 and 2005". In: *ECMWF Newsl.* 104, pp. 10–14 (cit. on p. 48).

[40] Candille, G. and Talagrand, O. (2005). "Evaluation of probabilistic prediction systems for a scalar variable". In: *Q. J. R. Meteorolog. Soc.* 131.609, pp. 2131–2150 (cit. on p. 29).

[41] Caruana, R., Niculescu-Mizil, A., Crew, G., and Ksikes, A. (2004). *Ensemble selection from libraries of models*. In: *Proceedings of the twenty-first international conference on Machine learning*, pp. 137–144 (cit. on p. 127).

[42] Chaudhari, P., Dharaskar, R., and Thakare, V. M. (2010). "Computing the Most Significant Solution from Pareto Front obtained in Multi-objective Evolutionary". In: *IJACSA* 1.4, pp. 63–68 (cit. on p. 92).

[43] Chen, C.-S., Chen, B. P.-T., Chou, F. N.-F., and Yang, C.-C. (2010). "Development and application of a decision group Back-Propagation Neural Network for flood forecasting". In: *J. Hydrol.* 385.1–4, pp. 173–182 (cit. on p. 10).

[44] Chryssolouris, G., Lee, M., and Ramsey, A. (1996). "Confidence interval prediction for neural network models". In: *IEEE Trans. on Neural Netw.* 7.1, pp. 229–232 (cit. on p. 11).

[45] Cloke, H. L. and Pappenberger, F. (2009). "Ensemble flood forecasting: A review". In: *J. Hydrol.* 375.3-4, pp. 613–626 (cit. on pp. 9, 23, 46, 47).

[46] Confesor, R. B. and Whittaker, G. W. (2007). "Automatic calibration of hydrologic models with multi-objective evolutionary algorithm and pareto optimization". In: *J. Am. Water Resour. Assoc.* 43, pp. 981–989 (cit. on p. 47).

[47] Corne, D. W. and Knowles, J. D. (2003). *No free lunch and free leftovers theorems for multiobjective optimisation problems*. In: *Proceedings of the 2nd international conference on Evolutionary multi-criterion optimization*. Springer-Verlag, pp. 327–341 (cit. on p. 126).

[48] Coulibaly, P. (2010). "Reservoir Computing approach to Great Lakes water level forecasting". In: *J. Hydrol.* 381.1–2, pp. 76–88 (cit. on p. 113).

[49] Coulibaly, P. and Baldwin, C. K. (2005). "Nonstationary hydrological time series forecasting using nonlinear dynamic methods". In: *J. Hydrol.* 307.1–4, pp. 164–174 (cit. on p. 113).

[50] Deb, K., Pratap, A., Agarwal, S., and Meyarivan, T. (2002). "A fast and elitist multiobjective genetic algorithm: NSGA-II". In: *IEEE Trans. Evol. Comp.* 6.2, pp. 182–197 (cit. on pp. 40, 91).

[51] Diamantidis, N., Karlis, D., and Giakoumakis, E. (2000). "Unsupervised stratification of cross-validation for accuracy estimation". In: *Artif. Intell.* 116.1-2, pp. 1–16 (cit. on pp. 34, 59, 107–109, 124).

[52] Domingos, P. (2012). "A few useful things to know about machine learning". In: *Commun. ACM* 55.10, pp. 78–87 (cit. on pp. 33, 107).

[53] Duan, Q., Schaake, J., Andréassian, V., Franks, S., Goteti, G., Gupta, H., Gusev, Y., Habets, F., Hall, A., Hay, L., Hogue, T., Huang, M., Leavesley, G., Liang, X., Nasonova, O., Noilhan, J., Oudin, L., Sorooshian, S., Wagener, T., and Wood, E. (2006). "Model Parameter Estimation Experiment (MOPEX): An overview of science strategy and major results from the second and third workshops". In: *J. Hydrol.* 320.1–2, pp. 3–17 (cit. on pp. 116, 118, 127, 144).

[54] Duan, Q., Ajami, N. K., Gao, X., and Sorooshian, S. (2007). "Multi-model ensemble hydrologic prediction using Bayesian model averaging". In: *Adv. Water Res.* 30.5, pp. 1371–1386 (cit. on p. 126).

[55] Duan, Q., Gupta, V., and Sorooshian, S. (1993). "Shuffled complex evolution approach for effective and efficient global minimization". In: *J. Optimiz. Theory App.* 76.3, pp. 501–521 (cit. on p. 2).

[56] Duda, R. O., Hart, P. E., and Stork, D. G. (2001). *Pattern Classification*. 2nd ed. Wiley Interscience (cit. on pp. 32, 95).

[57] Ebert, C., Bárdossy, A., and Bliefernicht, J. (2007). *Selecting members of an EPS for flood forecasting systems by using atmospheric circulation patterns*. In: *European Geosciences Union (EGU), Geophysical Research Abstract.* Vol. 9 (cit. on pp. 46, 84).

[58] Eiben, A. and Smith, J. (2003). *Introduction to evolutionary computing*. 1, Corr. 2nd printing. Springer, p. 299 (cit. on pp. 2, 38, 40, 91, 92).

[59] Farnsworth, R. K., Thompson, E. S., and L., P. E. (1982). *Evaporation Atlas for the Contiguous 48 United States*. Tech. rep. NOAA Technical Report NWS 33. National Oceanic and Atmospheric Administration, National Weather Service (cit. on p. 116).

[60] Ferranti, L. and Corti, S. (2011). "New clustering products". In: *ECMWF Newsl.* 127, pp. 6–12 (cit. on p. 84).

[61] Fleuret, F. (2004). "Fast Binary Feature Selection with Conditional Mutual Information". In: *JMLR* 5, pp. 1531–1555 (cit. on p. 89).

[62] Fortin, F.-A., Rainville, F.-M. D., Gardner, M.-A., Parizeau, M., and Gagné, C. (2012). "DEAP: Evolutionary Algorithms Made Easy". In: *JMLR* 13, pp. 2171–2175 (cit. on p. 100).

[63] Freund, Y. and Schapire, R. E. (1996). *Experiments with a New Boosting Algorithm*. In: *Machine Learning: Proceedings of the Thirteenth International Conference*, pp. 148–156 (cit. on pp. 126, 139).

[64] Gaborit, E., Anctil, F., Fortin, V., and Pelletier, G. (2013). "On the reliability of spatially disaggregated global ensemble rainfall forecasts". In: *Hydrol. Processes* 27.1, pp. 45–56 (cit. on p. 9).

[65] Geman, S., Bienenstock, E., and Doursat, R. (1992). "Neural networks and the bias variance dilemma". In: *Neural Comput.* 4.1, pp. 1–58 (cit. on pp. 21, 32).

[66] Gneiting, T. and Raftery, A. E. (2007). "Strictly proper scoring rules, prediction, and estimation". In: *J. Am. Statist. Assoc.* 102.477, pp. 359–378 (cit. on pp. 19, 26, 27, 59, 76).

[67] Good, I. J. (1952). "Rational Decisions". In: *J. R. Stat. Soc.* 14.1, pp. 107–114 (cit. on p. 26).

[68] Gouweleeuw, B. T., Thielen, J., Franchello, G., De Roo, A. P. J., and Buizza, R. (2005). "Flood forecasting using medium-range probabilistic weather prediction". In: *Hydrol. Earth Syst. Sci.* 9.4, pp. 365–380 (cit. on p. 47).

[69] Gupta, H. V., Bastidas, L. A., Sorooshian, S., Shuttleworth, W. J., and Yang, Z. L. (1999). "Parameter estimation of a land surface scheme using multicriteria methods". In: *J. Geophys. Res.* 104.D16, pp. 19491–19503 (cit. on p. 47).

[70] Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F. (2009). "Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling". In: *J. Hydrol.* 377.1-2, pp. 80–91 (cit. on p. 20).

[71] Gupta, H. V., Sorooshian, S., and Yapo, P. O. (1998). "Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information". In: *Water Resour. Res.* 34.4, pp. 751–763 (cit. on p. 47).

[72] Gupta, V. K. and Sorooshian, S. (1985). "The relationship between data and the precision of parameter estimates of hydrologic models". In: *J. Hydrol.* 81.1-2, pp. 57–77 (cit. on p. 107).

[73] Guyon, I. and Elisseeff, A. (2003). "An introduction to variable and feature selection". In: *JMLR* 3, pp. 1157–1182 (cit. on p. 38).

[74] Hagan, M. T., Demuth, H. B., and Beale, M. (1996). *Neural network design*. 1st ed. PWS Publishing Co., p. 730 (cit. on pp. 34, 37, 115, 146).

[75] Haindl, M., Somol, P., Ververidis, D., and Kotropoulos, C. (2006). "Feature Selection Based on Mutual Correlation". In: *Progress in Pattern Recognition, Image Analysis and Applications*. Vol. 4225. Springer Berlin Heidelberg, pp. 569–577 (cit. on pp. 87, 88).

[76] Hamill, T. M. and Colucci, S. J. (1997). "Verification of Eta–RSM Short-Range Ensemble Forecasts". In: *Mon. Weather Rev.* 125.6, pp. 1312–1327 (cit. on pp. 29, 56).

[77] Haykin, S. (2001). *Feedforward neural networks: An introduction*. In: Nonlinear Dynamical Systems: Feedforward Neural Network Perspectives. Chap. 1, pp. 1–16 (cit. on pp. 34, 131).

[78] He, Y., Wetterhall, F., Cloke, H. L., Pappenberger, F., Wilson, M., Freer, J., and McGregor, G. (2009). "Tracking the uncertainty in flood alerts driven by grand ensemble weather predictions". In: *Meteorol. Appl.* 16.1, pp. 91–101 (cit. on pp. 9, 46).

[79] Hersbach, H. (2000). "Decomposition of the Continuous Ranked Probability Score for ensemble prediction systems". In: *Wea. Forecasting* 15.5, pp. 559–570 (cit. on pp. 25, 26).

[80] Hettiarachchi, P., Hall, M. J., and Minns, A. W. (2005). "The extrapolation of artificial neural networks for the modelling of rainfall–runoff relationships". In: *J. Hydroinf.* 7.4, pp. 291–296 (cit. on pp. 106, 119).

[81] Hornik, K., Stinchcombe, M., and White, H. (1989). "Multilayer feedforward networks are universal approximators". In: *Neural Networks* 2.5, pp. 359–366 (cit. on p. 37).

[82] Hwang, J. T. G. and Ding, A. A. (1997). "Prediction Intervals for Artificial Neural Networks". In: *J. Am. Statist. Assoc.* 92.438, pp. 748–757 (cit. on p. 11).

[83] Jaun, S., Ahrens, B., Walser, A., Ewen, T., and Schär, C. (2008). "A probabilistic view on the August 2005 floods in the upper Rhine catchment". In: *Nat. Hazards Earth Syst. Sci.* 8, pp. 281–291 (cit. on pp. 3, 46, 84, 93).

[84] Kasiviswanathan, K. and Sudheer, K. (2013). "Quantification of the predictive uncertainty of artificial neural network based river flow forecast models". In: *Stoch. Environ. Res. Risk Assess.* 27.1, pp. 137–146 (cit. on p. 11).

[85] Khosravi, A., Nahavandi, S., Creighton, D., and Atiya, A. (2011). "Comprehensive Review of Neural Network-Based Prediction Intervals and New Advances". In: *IEEE Trans. on Neural Netw.* 22.9, pp. 1341–1356 (cit. on p. 11).

[86] Kirchner, J. W. (2006). "Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology". In: *Water Resour. Res.* 42, pp. 1–5 (cit. on pp. 2, 32).

[87] Kistler, R., Collins, W., Saha, S., White, G., Woollen John ans Kalnay, E., Chelliah, M., Ebisuzaki, W., Kanamitsu, M., Kousky, V., Dool, H. van den, Jenne, R., and Fiorino, M. (2001). "The NCEP–NCAR 50–Year Reanalysis: Monthly Means CD–ROM and Documentation". In: *Bull. Amer. Meteor. Soc.* 82, pp. 247–267 (cit. on p. 118).

[88] Kohavi, R. and John, G. H. (1997). "Wrappers for feature subset selection". In: *Artif. Intell.* 97.1–2, pp. 273–324 (cit. on p. 38).

[89] Kottegoda, N. T. and Rosso, R. (2009). *Applied Statistics for Civil and Environmental Engineers*. Wiley, p. 737 (cit. on pp. 25, 86).

[90] Krogh, A. and Vedelsby, J. (1995). *Neural Network Ensembles, Cross Validation, and Active Learning*. In: *Advances in Neural Information Processing Systems 8*. MIT Press, pp. 231–238 (cit. on pp. 11, 21).

[91] Kuczera, G. (1982). "On the relationship between the reliability of parameter estimates and hydrologic time series data used in calibration". In: *Water Resour. Res.* 18.1, pp. 146–154 (cit. on p. 107).

[92] Kuncheva, L. I. (2004). *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience, p. 350 (cit. on pp. 6, 11, 82, 126, 140).

[93] Kunstmann, H., Jung, G., Wagner, S., and Clottey, H. (2008). "Integration of atmospheric sciences and hydrology for the development of decision support systems in sustainable water management". In: *Phys. Chem. Earth* 33.1–2, pp. 165–174 (cit. on p. 2).

[94] Kwak, N. and Choi, C.-H. (2002). "Input feature selection for classification problems". In: *IEEE Trans. on Neural Netw.* 13.1, pp. 143 –159 (cit. on p. 89).

[95] Liu, Y., Weerts, A. H., Clark, M., Hendricks Franssen, H.-J., Kumar, S., Moradkhani, H., Seo, D.-J., Schwanenberg, D., Smith, P., Dijk, A. I. J. M. van, Velzen, N. van, He, M., Lee, H., Noh, S. J., Rakovec, O., and Restrepo, P. (2012). "Advancing data assimilation in operational hydrologic forecasting: progresses, challenges, and emerging opportunities". In: *Hydrol. Earth Syst. Sci.* 16.10, pp. 3863–3887 (cit. on p. 9).

[96] López, K., Gagné, C., Castellanos, G., and Orozco, M. (2013). *Training Subset Selection in Hourly Ontario Energy Price Forecasting using Time Series Clustering-based Stratification*. In preparation (cit. on pp. 107, 108, 123).

[97] Maier, H. R. and Dandy, G. C. (2000). "Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications". In: *Environ. Modell. Softw.* 15.1, pp. 101–124 (cit. on pp. 3, 9, 14, 37, 107, 113, 140).

[98] Maier, H. R., Jain, A., Dandy, G. C., and Sudheer, K. (2010). "Methods used for the development of neural networks for the prediction of water resource variables in river systems: Current status and future directions". In: *Environ. Modell. Softw.* 25.8, pp. 891–909 (cit. on pp. 9, 14, 37, 113, 140).

[99] Marler, R. and Arora, J. (2004). "Survey of multi-objective optimization methods for engineering". In: *Struct. Multidiscip. O.* 26 (6), pp. 369–395 (cit. on pp. 87, 100).

[100] Marsigli, C., Montani, A., Nerozzi, F., Paccagnella, T., Tibaldi, S., Molteni, F., and Buizza, R. (2001). "A strategy for high-resolution ensemble prediction. II: Limited-area experiments in four Alpine flood events". In: *Q. J. R. Meteorolog. Soc.* 127, pp. 2095–2115 (cit. on pp. 46, 84).

[101] Martinez Alvarez, F., Troncoso, A., Riquelme, J., and Aguilar Ruiz, J. (2011). "Energy Time Series Forecasting Based on Pattern Sequence Similarity". In: *IEEE Trans. Knowl. Data Eng.* 23.8, pp. 1230–1243 (cit. on pp. 92, 108, 109).

[102] Marty, R., Zin, I., and Obled, C. (2013). "Sensitivity of hydrological ensemble forecasts to different sources and temporal resolutions of probabilistic quantitative precipitation forecasts: flash flood case studies in the Cévennes-Vivarais region (Southern France)". In: *Hydrol. Processes* 27.1, pp. 33–44 (cit. on p. 9).

[103] May, R., Maier, H., and Dandy, G. (2010). "Data splitting for artificial neural networks using SOM-based stratified sampling". In: *Neural Networks* 23.2, pp. 283–294 (cit. on pp. 106, 107).

[104] May, R. J., Dandy, G. C., Maier, H. R., and Nixon, J. B. (2008). "Application of partial mutual information variable selection to ANN forecasting of water quality in water distribution systems". In: *Environ. Modell. Softw.* 23, pp. 1289 –1299 (cit. on p. 127).

[105] McKay, M. D., Beckman, R. J., and Conover, W. J. (1979). "A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code". In: *Technometrics* 21.2, pp. 239–245 (cit. on p. 91).

[106] Mitchell, T. (1997). *Machine Learning*. 1st ed. McGraw-Hill Education, p. 432 (cit. on pp. 33, 34).

[107] Moffet, R., Gagnon, N., and Fontecilla, J. (2010). *Monthly and seasonal weather - is it predictable?* In: *Manitoba Agronomists Conference*. University of Manitoba (cit. on p. 4).

[108] Molteni, F., Buizza, R., Palmer, T. N., and Petroliagis, T. (1996). "The ECMWF Ensemble Prediction System: Methodology and validation". In: *Q. J. R. Meteorolog. Soc.* 122, pp. 73–119 (cit. on p. 48).

[109] Molteni, F., Buizza, R., Marsigli, C., Montani, A., Nerozzi, F., and Paccagnella, T. (2001). "A strategy for high-resolution ensemble prediction. I: Definition of representative members and global-model experiments". In: *Q. J. R. Meteorolog. Soc.* 127, pp. 2069–2094 (cit. on pp. 46, 82, 84).

[110] Moore, D. S., McCabe, G. P., and Craig, B. A. (2009). *Introduction to the practice of statistics*. 6th ed. W. H. Freeman and Company, p. 709 (cit. on p. 21).

[111] Murphy, A. H. (1993). "What is a good forecast? An essay on the nature of goodness in weather forecasting". In: *Wea. Forecasting* 8, pp. 281–293 (cit. on pp. 5, 19, 20, 23, 126).

[112] Murphy, A. H. (1988). "Skill Scores Based on the Mean Square Error and Their Relationships to the Correlation Coefficient". In: *Mon. Weather Rev.* 116, pp. 2417–2424 (cit. on p. 20).

[113] Nguyen, D. and Widrow, B. (1990). *Improving the learning speed of 2-layer neural networks by choosing initial values of the adaptive weights*. In: *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*. Vol. 3. IEEE, pp. 21–26 (cit. on pp. 114, 115).

[114] Nix, D. and Weigend, A. (1994). *Estimating the mean and variance of the target probability distribution*. In: *IEEE International Conference on Neural Networks*. Vol. 1, pp. 55–60 (cit. on p. 11).

[115] Pappenberger, F., Beven, K. J., Hunter, N. M., Bates, P. D., Gouweleeuw, B. T., Thielen, J., and Roo, A. P. J. de (2005). "Cascading model uncertainty from medium range weather forecasts (10 days) through a rainfall-runoff model to flood inundation predictions within the European Flood Forecasting System (EFFS)". In: *Hydrol. Earth Syst. Sci.* 9.4, pp. 381–393 (cit. on pp. 3, 5, 9, 10, 46, 93, 126, 145).

[116] Parkes, B. L., Wetterhall, F., Pappenberger, F., He, Y., Malamud, B. D., and Cloke, H. (2013). "Assessment of a 1 hour gridded precipitation dataset to drive a hydrological model: a case study of the summer 2007 floods in the Upper Severn, UK". In: *Hydrol. Res.* 44.1, pp. 89–105 (cit. on p. 9).

[117] Peng, H., Long, F., and Ding, C. (2005). "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 27.8, pp. 1226–1238 (cit. on p. 89).

[118] Persson, A. (2011). *User Guide to ECMWF forecast products*. European Centre for Medium Range Weather Forecasts (cit. on p. 84).

[119] Pietroniro, A., Fortin, V., Kouwen, N., Neal, C., Turcotte, R., Davison, B., Verseghy, D., Soulis, E. D., Caldwell, R., Evora, N., and Pellerin, P. (2007). "Development of the MESH modelling system for hydrological ensemble forecasting of the Laurentian Great Lakes at the regional scale". In: *Hydrol. Earth Syst. Sci.* 11.4, pp. 1279–1294 (cit. on p. 2).

[120] Pilgrim, D. H., Chapman, T. G., and Doran, D. G. (1988). "Problems of rainfall-runoff modelling in arid and semiarid regions". In: *Hydrol. Sci. J.* 33.4, pp. 379–400 (cit. on pp. 116, 120).

[121] Ponce, V., Pandey, R., and Ercan, S. (2000). "Characterization of Drought across Climatic Spectrum". In: *J. Hydrol. Eng.* 5.2, pp. 222–224 (cit. on p. 116).

[122] Quintana-Seguí, P., Le Moigne, P., Durand, Y., Martin, E., Habets, F., Baillon, M., Canellas, C., Franchisteguy, L., and Morel, S. (2008). "Analysis of Near-Surface Atmospheric Variables: Validation of the SAFRAN Analysis over France". In: *J. Appl. Meteor. Climatol.* 47.1, pp. 92–107 (cit. on p. 48).

[123] Raftery, A. E., Gneiting, T., Balabdaoui, F., and Polakowski, M. (2005). "Using Bayesian Model Averaging to calibrate forecast ensembles". In: *Mon. Weather Rev.* 133.5, pp. 1155–1174 (cit. on p. 46).

[124] Ramos, M. H., Andel, S. J. van, and Pappenberger, F. (2013). "Do probabilistic forecasts lead to better decisions?" In: *Hydrol. Earth Syst. Sci.* 17.6, pp. 2219–2232 (cit. on pp. 7, 8).

[125] Renner, M., Werner, M., Rademacher, S., and Sprokkereef, E. (2009). "Verification of ensemble flow forecasts for the River Rhine". In: *J. Hydrol.* 376.3-4, pp. 463–475 (cit. on p. 47).

[126] Rojas, R. (1996). *Neural Networks: A Systematic Introduction*. 1st ed. Springer (cit. on p. 37).

[127] Roulin, E. (2007). "Skill and relative economic value of medium-range hydrological ensemble predictions". In: *Hydrol. Earth Syst. Sci.* 11.2, pp. 725–737 (cit. on pp. 45, 126).

[128] Roulston, M. S. and Smith, L. A. (2002). "Evaluating Probabilistic Forecasts Using Information Theory". In: *Mon. Weather Rev.* 130.6, pp. 1653–1660 (cit. on p. 27).

[129] Rousset, F., Habets, F., Martin, E., and Noilhan, J. (2007). "Ensemble streamflow forecasts over France". In: *ECMWF Newsl.* 111, pp. 21–27 (cit. on pp. 45, 47).

[130] Schaake, J. C., Hamill, T. M., Buizza, R., and Clark, M. (2007). "HEPEX: The Hydrological Ensemble Prediction Experiment". In: *Bull. Am. Meteorol. Soc.* 88.10, pp. 1541–1547 (cit. on p. 4).

[131] Schaake, J., Pailleux, J., Thielen, J., Arritt, R., Hamill, T., Luo, L., Martin, E., McCollor, D., and Pappenberger, F. (2010). "Summary of recommendations of the first workshop on Postprocessing and Downscaling Atmospheric Forecasts for Hydrologic Applications held at Météo-France, Toulouse, France, 15–18 June 2009". In: *Atmos. Sci. Lett.* 11.2, pp. 59–63 (cit. on p. 9).

[132] See, L. and Abrahart, R. J. (2001). "Multi-model data fusion for hydrological forecasting". In: *Computers & Geosciences* 27.8, pp. 987–994 (cit. on p. 126).

[133] Senbeta, D., Shamseldin, A., and O'Connor, K. (1999). "Modification of the probability-distributed interacting storage capacity model". In: *J. Hydrol.* 224.3–4, pp. 149–168 (cit. on p. 131).

[134] Shahin, M., Maier, H., and Jaksa, M. (2004). "Data Division for Developing Neural Networks Applied to Geotechnical Engineering". In: *J. Comput. Civ. Eng.* 18.2, pp. 105–114 (cit. on p. 106).

[135] Shao, Y., Taff, G., and Walsh, S. (2011). "Comparison of Early Stopping Criteria for Neural-Network-Based Subpixel Classification". In: *IEEE Geosci. Remote Sens. Lett.* 8.1, pp. 113–117 (cit. on p. 37).

[136] Shrestha, D. L. and Solomatine, D. P. (2006). "Experiments with AdaBoost.RT, an improved boosting scheme for regression". In: *Neural Comput.* 18.7, pp. 1678–1710 (cit. on pp. 6, 11, 126, 139).

[137] Silverman, B. W. (1986). *Density estimation: for statistics and data analysis* (cit. on p. 86).

[138] Sivapalan, M., Zhang, L., Vertessy, R., and Blöschl, G. (2003). "Downward approach to hydrological prediction". In: *Hydrol. Processes* 17.11, pp. 2099–2099 (cit. on pp. 3, 32).

[139] Solomatine, D. and Shrestha, D. (2004). *AdaBoost.RT: a boosting algorithm for regression problems*. In: *IEEE International Joint Conference on Neural Networks*. Vol. 2, pp. 1163–1168 (cit. on p. 6).

[140] Sorooshian, S. and Gupta, V. K. (1983). "Automatic calibration of conceptual rainfall-runoff models: The question of parameter observability and uniqueness". In: *Water Resour. Res.* 19.1, pp. 260–268 (cit. on p. 107).

[141] Székely, G. J. (2000). *$\mathcal{E}$-Statistics: The energy of statistical samples*. Tech. rep. 03-05. Department of Mathematics and Statistics, Bowling Green State University (cit. on p. 25).

[142] Talagrand, O., Vautard, R., and Strauss, B. (1997). *Evaluation of probabilistic prediction systems*. In: *Workshop on predictability*, pp. 1–25 (cit. on p. 29).

[143] Taylor, K. E. (2001). "Summarizing multiple aspects of model performance in a single diagram". In: *J. Geophys. Res.* 106.D7, pp. 7183–7192 (cit. on p. 55).

[144] Thirel, G., Rousset-Regimbeau, F., Martin, E., and Habets, F. (2008). "On the Impact of Short-Range Meteorological Forecasts for Ensemble Streamflow Predictions". In: *J. Hydrometeor.* 9.6, pp. 1301–1317 (cit. on p. 47).

[145] Todini, E. (2004). "Role and treatment of uncertainty in real-time flood forecasting". In: *Hydrol. Processes* 18.14, pp. 2743–2746 (cit. on pp. 3, 46, 93).

[146] Ueda, N. and Nakano, R. (1996). *Generalization error of ensemble estimators*. In: *IEEE International Conference on Neural Networks*. Vol. 1, pp. 90–95 (cit. on p. 22).

[147] Vafaie, H. and De Jong, K. (1992). *Genetic algorithms as a tool for feature selection in machine learning*. In: *Fourth International Conference on Tools with Artificial Intelligence*, pp. 200–203 (cit. on p. 38).

[148] Velázquez, J. A., Anctil, F., Ramos, M. H., and Perrin, C. (2011). "Can a multi-model approach improve hydrological ensemble forecasting? A study on 29 French catchments using 16 hydrological model structures". In: *Adv. Geosci.* 29, pp. 33–42 (cit. on pp. 5, 9, 10, 12, 45, 48, 50–52, 56, 59, 69, 76, 93, 127).

[149] Vos, N. J. de (2013). "Echo state networks as an alternative to traditional artificial neural networks in rainfall–runoff modelling". In: *Hydrol. Earth Syst. Sci.* 17.1, pp. 253–267 (cit. on pp. 10, 14, 113, 118, 123, 138, 140, 144).

[150] Vrugt, J. A. and Robinson, B. A. (2007). *Improved evolutionary optimization from genetically adaptive multimethod search*. In: *Proceedings of the National Academy of Sciences of the United States of America*. Vol. 104. 3, pp. 708–711 (cit. on p. 47).

[151] Vrugt, J., Robinson, B., and Hyman, J. (2009). "Self-Adaptive Multimethod Search for Global Optimization in Real-Parameter Spaces". In: *IEEE Trans. Evol. Comp.* 13.2, pp. 243–259 (cit. on p. 91).

[152] Vrugt, J., Diks, C., and Clark, M. (2008). "Ensemble Bayesian model averaging using Markov Chain Monte Carlo sampling". In: *Environ. Fluid Mech.* 8 (5), pp. 579–595 (cit. on pp. 25, 46).

[153] Vrugt, J. A., Gupta, H. V., Bouten, W., and Sorooshian, S. (2003a). "A Shuffled Complex Evolution Metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters". In: *Water Resour. Res.* 39.8, pp. 1–14 (cit. on p. 2).

[154] Vrugt, J. A., Gupta, H. V., Bastidas, L. A., Bouten, W., and Sorooshian, S. (2003b). "Effective and efficient algorithm for multiobjective optimization of hydrologic models". In: *Water Resour. Res.* 39.8, pp. 1–19 (cit. on p. 2).

[155] Vrugt, J. A., Clark, M. P., Diks, C. G. H., Duan, Q., and Robinson, B. A. (2006). "Multi-objective calibration of forecast ensembles using Bayesian model averaging". In: *Geophys. Res. Lett.* 33.19, pp. 1–6 (cit. on pp. 19, 46, 47).

[156] Wagener, T., Boyle, D. P., Lees, M. J., Wheater, H. S., Gupta, H. V., and Sorooshian, S. (2001). "A framework for development and application of hydrological models". In: *Hydrol. Earth Syst. Sci.* 5.1, pp. 13–26 (cit. on p. 47).

[157] Weigend, A. S. and Shi, S. (2000). "Predicting daily probability distributions of S&P500 returns". In: *J. Forecast.* 19.4, pp. 375–392 (cit. on p. 27).

[158] Wetterhall, F., Pappenberger, F., Cloke, H. L., Pozo, J. Thielen-del, Balabanova, S., Daňhelka, J., Vogelbacher, A., Salamon, P., Carrasco, I., Cabrera-Tordera, A. J., Corzo-Toscano, M., Garcia-Padilla, M., Garcia-Sanchez, R. J., Ardilouze, C., Jurela, S., Terek, B., Csik, A., Casey, J., Stanǔnavičius, G., Ceres, V., Sprokkereef, E., Stam, J., Anghel, E., Vladikovic, D., Alionte Eklund, C., Hjerdt, N., Djerv, H., Holmberg, F., Nilsson, J., Nyström, K., Sušnik, M., Hazlinger, M., and Holubecka, M. (2013). "Forecasters priorities for improving probabilistic flood forecasts". In: *Hydrol. Earth Syst. Sci. Discuss.* 10.2, pp. 2215–2242 (cit. on pp. 5, 8).

[159] Whitley, D. (2000). *Evolutionary Computation 1: Basic Algorithms and Operators*. In: Institute of Physics Publishing. Chap. 33.3, pp. 274–284 (cit. on p. 92).

[160] Wilks, D. S. (2011a). "On the Reliability of the Rank Histogram". In: *Mon. Weather Rev.* 139.1, pp. 311–316 (cit. on pp. 30, 62).

[161] Wilks, D. S. (2011b). *Statistical Methods in the Atmospheric Sciences*. 3rd ed. Vol. 100. Academic Press, p. 676 (cit. on pp. 23, 25, 28, 47, 63).

[162] Williams, G. P. (1997). *Chaos Theory Tamed*. 1st ed. National Academies Press, p. 406 (cit. on p. 27).

[163] Wolpert, D. H. (1992). "Stacked generalization". In: *Neural Networks* 5.2, pp. 241–259 (cit. on p. 126).

[164] Xuan, Y., Cluckie, I. D., and Wang, Y. (2009). "Uncertainty analysis of hydrological ensemble forecasts in a distributed model utilising short-range rainfall prediction". In: *Hydrol. Earth Syst. Sci.* 13.3, pp. 293–303 (cit. on pp. 46, 84).

[165] Yang, H. H. and Moody, J. (1999). *Data Visualization and Feature Selection: New Algorithms for Nongaussian Data*. In: *Advances In Neural Information Processing Systems 12*. MIT Press, pp. 687–693 (cit. on p. 89).

[166] Yang, J. and Honavar, V. (1998). "Feature subset selection using a genetic algorithm". In: *IEEE Intell. Syst.* 13.2, pp. 44–49 (cit. on p. 38).

[167] Yapo, P. O., Gupta, H. V., and Sorooshian, S. (1998). "Multi-objective global optimization for hydrologic models". In: *J. Hydrol.* 204.1-4, pp. 83–97 (cit. on p. 47).

[168] Yapo, P. O., Gupta, H. V., and Sorooshian, S. (1996). "Automatic calibration of conceptual rainfall-runoff models: sensitivity to calibration data". In: *J. Hydrol.* 181.1–4, pp. 23–48 (cit. on p. 107).

[169] Zabel, F. and Mauser, W. (2013). "2-way coupling the hydrological land surface model PROMET with the regional climate model MM5". In: *Hydrol. Earth Syst. Sci.* 17.5, pp. 1705–1714 (cit. on p. 2).