Proceedings of the 19th World Congress
The International Federation of Automatic Control
Cape Town, South Africa. August 24-29, 2014

Universal approximators for direct policy search in multi-purpose water reservoir management: A comparative analysis

Matteo Giuliani * **Emanuele Mason** * **Andrea Castelletti** *
Francesca Pianosi ** **Rodolfo Soncini-Sessa** *

* *Department of Electronics, Information, and Bioengineering, Politecnico di Milano, Milan, Italy (e-mail: matteo.giuliani@polimi.it; emanuele.mason@polimi.it; andrea.castelletti@polimi.it; rodolfo.soncini@polimi.it).*
** *Department of Civil Engineering, University of Bristol, Bristol, UK (e-mail: francesca.pianosi@bristol.ac.uk).*

Abstract: This study presents a novel approach which combines direct policy search and multi-objective evolutionary algorithms to solve high-dimensional state and control space water resources problems involving multiple, conflicting, and non-commensurable objectives. In such a multi-objective context, the use of universal function approximators is generally suggested to provide flexibility to the shape of the control policy. In this paper, we comparatively analyze Artificial Neural Networks (ANN) and Radial Basis Functions (RBF) under different sets of input to estimate their scalability to high-dimensional state space problems. The multi-purpose HoaBinh water reservoir in Vietnam, accounting for hydropower production and flood control, is used as a case study. Results show that the RBF policy parametrization is more effective than the ANN one. In particular, the approximated Pareto front obtained with RBF control policies successfully explores the full tradeoff space between the two conflicting objectives, while the ANN solutions are often Pareto-dominated by the RBF ones.

Keywords: radial basis functions; artificial neural networks; direct policy search; multi-objective optimization; environmental engineering

1. INTRODUCTION

The optimal operation of water resources systems is a wide and challenging application domain for optimal control methodologies and tools. Most of the operation problems involving water resources systems can be formulated as Markov decision processes (MDP, see White (1982)) and solved via Dynamic Programming or Reinforcement Learning (Powell, 2007; Busoniu et al., 2010). For example, water reservoir operations are a sequence of release decisions, made at discrete time instants, over a system affected by stochastic disturbances (i.e., inflows). Similarly, well field operations are a sequence of pumping rate decisions, made at discrete time instants and different points in space, over a system affected by stochastic disturbance (i.e., groundwater recharge).

Although DP family methods can be applied under mild assumptions, such as the disturbance process is time-independent and the objective functions must be time-separable, they suffer from a well known dual curse which prevents them from being employed to solve large-scale control schemes: *i*) the curse of dimensionality (Bellman, 1957), namely the computational cost of DP grows exponentially with state, decision, and disturbance vectors; and *ii*) the curse of modeling (Tsitsiklis and Van Roy, 1996), meaning the use of in-line model-based computations that make impossible the direct, model-free use of exogenous information into the controller and the use of process-based simulation models (e.g., hydrodynamic and ecological). Yet, when adapted to the water resources domain, DP-based approaches suffer from another curse that might overly affect the computational requirements: DP methods are inherently single-objective, while the management of water resources very often involves multiple, incommensurable objectives. The only way to turn DP into a multi-objective algorithm is to re-formulate the multi-objective control problem as a family of parametric single-objective problems and reiteratively running a single-objective optimization for different values of the parameters to approximate the continuous Pareto front of the original multi-objective problem. This remarkably affects the computational requirements, as the number of single-objective problems to solve grows exponentially with the objectives number and most of the water resources problems involves more than four or five objectives (e.g., Castelletti et al., 2013a).

Approximate DP methods have been developed to overcome the above limitations. However, most of them (e.g., model predictive control, chance-constrained control) maintains the DP structure, computing approxima-

⋆ Matteo Giuliani was partially supported by *Fondazione Fratelli Confalonieri*. Emanuele Mason was supported by the *Integrated and sustainable water Management of the Red-Thai Binh Rivers System in changing climate.*

tions of the Bellman function and solving single-objective problems. In this paper, instead, we combine the direct policy search (DPS, see Rosenstein and Barto (2001)) with multi-objective evolutionary algorithms (MOEAs), in order to solve high-dimension state and control space problems and find an approximation of the entire Pareto front, and the associated control policies, in a single optimization run. DPS, also known as parameterization-simulation-optimization in the water resources literature (Koutsoyiannis and Economou, 2003), is a simulation-based approach where the control policy is first parameterized within a given family of functions and then the parameters optimized with respect to the objectives of the control problem. The selection of a suitable class of functions to which the control policy belong to is a key operation, as it might restrict the search for the optimal policy to a subspace of the decision space that does not include the optimal solution. In the water reservoir literature, a number of parameterizations have been proposed (see Lund and Guzman, 1999, and references therein). However, they are based largely on empirical or experimental successes and are designed, mostly via simulation, for single-purpose reservoirs (Lund and Guzman, 1999). In a multi-objective context similar rules can not easily inferred from the experience and the use of universal function approximators (Tikk et al., 2003) is generally preferred. In principle, universal approximators should be capable of accurately estimating any unknown continuous function under very mild assumptions. In practice, the optimal training of the approximator, and thus its accuracy, strongly depends on the parameters domain, on the dimensions of the input (state) and output (control) sets, and on the size of the training dataset available (Kurková and Sanguineti, 2001).

In this paper, we comparatively analyze two among the most commons universal approximators: Artificial Neural Networks (Hornik et al., 1989) and Radial Basis Functions (Liao et al., 2003). In particular, we assess their effectiveness under different sets of input to estimate their scalability to high-dimensional state space problems. The multi-objective optimization of the policy parameters is performed using the self-adaptive Borg MOEA (Hadka and Reed, 2013), which has been shown to be highly robust across a diverse suite of challenging multi-objective problems, where it met or exceeded the performance of other state-of-the-art MOEAs (Hadka and Reed, 2012). As a study site for the analysis we used the HoaBinh water reservoir system, a multi-purpose regulated reservoir in the Red River basin (Vietnam). The Red River Basin is the second largest basin of Vietnam and the HoaBinh reservoir is regulated to maximize hydropower production and flood control in Hanoi.

2. DIRECT POLICY SEARCH

Water reservoir operation problems generally require to take sequential decisions $\mathbf{u}_t$ at discrete time instants ($t = 1, 2, \ldots$) on the basis of the current system conditions described by the state vector $\mathbf{x}_t \in \mathbb{R}^{n_x}$. The controls $\mathbf{u}_t \in \mathcal{U}_t(\mathbf{x}_t) \subseteq \mathbb{R}^{n_u}$ are determined by a control policy π and alter the state of the system according to a probabilistic transition function $p(\mathbf{x}_{t+1}|\mathbf{x}_t, \mathbf{u}_t)$. The combination of states and controls defines a trajectory τ over the horizon H, which allows the evaluation of the policy π as follows:

$$J_\pi = \mathrm{E}[R(\tau)|\pi] = \int R(\tau) p_\pi(\tau) d\tau \tag{1}$$

where $R(\tau)$ defines the objective function of the problem and $p_\pi(\tau)$ is the distribution over trajectories τ. DP family methods estimate the expected long-term cost of a policy for each state $\mathbf{x}_t$ at time t by means of the value function $V_t^\pi(\mathbf{x}_t)$, which is defined over a discrete grid of time and state variables. The optimal policy π^* is then derived as the one maximizing the value function.

Direct policy search (DPS, see Rosenstein and Barto (2001)) directly operates in the policy space and avoids the computation of the value function. DPS is based on the parameterization of the policies π_θ and the exploration of the parameter space Θ with the aim to find a parameterized policy that optimizes the expected long-term performance (assumed to be a cost), i.e.

$$\pi_\theta^* = \arg\min_{\pi_\theta} J_\theta \tag{2}$$

where the policy π_θ is parameterized by parameters $\theta \in \Theta$. Finding π_θ^* is equivalent to find the corresponding optimal policy parameters θ^*. Problem (2) is dynamically constrained by the transition function of the system $p(\mathbf{x}_{t+1}|\mathbf{x}_t, \mathbf{u}_t)$. However, DPS does not provide any theoretical guarantee on the optimality of the resulting operating policies, which are strongly dependent on the choice of the class of functions to which they belong (Section 2.1) and on the ability of the optimization algorithm to deal with non-linear models and objectives functions, complex and highly constrained decision spaces, and many conflicting objectives (Section 2.2).

Different DPS approaches have been proposed in the last decades and they differ in the methods used for the generation of the trajectories as well as for the update and evaluation of the policies (for a review, see Deisenroth et al., 2011, and references therein). In this paper we use a DPS method with the following features:

- *Model-based approach*: dealing with natural systems, learning a policy through experiments on the real system is not possible, as it is time consuming and might require to acquire experience in dangerous states of the system. A dynamic model replaces the real system to perform simulations, based on which the policy is determined.
- *Stochastic trajectory generation*: the dynamic model of the system is used as simulator for sampling the trajectories τ. The system evolves according to the probabilistic transition function $p(\mathbf{x}_{t+1}|\mathbf{x}_t, \mathbf{u}_t)$ due to the presence of stochastic disturbances (e.g., reservoir inflow). In particular, we assume the average value over the time series is equivalent to the expected value over the probability distribution of the disturbances (Pianosi et al., 2011). The value of the objective function is approximated via simulation over a sufficiently long historical or synthetically generated realization of the disturbances. Deterministic trajectory prediction, instead, does not sample the simulated trajectories but analytically predicts the trajectories distribution $p_\theta(\tau)$.
- *Episode-based exploration and evaluation*: the quality of a parameter vector θ is evaluated as the expected return computed on the whole episode, with the parameter vector θ that changes at the start of the episode. Conversely, step-based exploration and

Algorithm 1 Stochastic model-based, episode-based, multi-objective DPS.

Initialization:
Random generation of a population $\{\theta^1, \dots, \theta^P\}$
Iterations: repeat until stopping conditions are met
- generation of the trajectory τ^i via model simulation according to the probabilistic transition function $p(\mathbf{x}_{t+1}|\mathbf{x}_t, \mathbf{u}_t)$ and following the policy π_{θ^i} (with $i = 1, \dots, P$)
- compute $J^1_{\theta^i}, \dots, J^q_{\theta^i}$, with $i = 1, \dots, P$
- generation of a new population by selection, crossover and mutation with respect to the best individuals (i.e., the solutions non Pareto-dominated)

evaluation assesses the quality of single state-control pairs changing the parameters at each time step, but it requires the evaluation of the value function. Moreover, episode-based methods are not restricted to time-separable cost functions, which can depend on the entire trajectory τ.

- *Multi-objective*: although most of DPS approaches relies on gradient-based single-objective optimization (e.g., Peters and Schaal, 2008), water reservoir problems are generally framed in complex socio-economic contexts requiring to consider multiple, non-commensurable operating objectives. The single-objective formulation (eqs. 1-2) is replaced by a vector of q objective functions $\mathbf{J} = [J^1, \dots, J^q]$. Multi-objective evolutionary algorithms (MOEAs) can be adopted in DPS problems to obtain an approximation of the Pareto front in a single run of the algorithm.

A tabular version of the stochastic, model-based, episode-based, multi-objective DPS approach is reported in Algorithm 1. In the next two sections the universal approximators and the MOEA algorithm considered in this study are described.

2.1 Universals approximators

In the literature, a number of parameterizations of water reservoir operating rules have been proposed. However, most of them are based on empirical or experimental successes and were designed, mostly via simulation, for single-purpose reservoirs (Lund and Guzman, 1999). In complex multi-objective problems, the adoption of universal approximators is generally preferred as it provides more flexibility to the shape of the control policy. In this paper, we define the parameterized policy π_θ using Artificial Neural Networks (e.g., Hornik et al., 1989) and gaussian Radial Basis Functions (e.g., Liao et al., 2003). The policy input vector χ_t includes the system state $\mathbf{x}_t$, with the resulting policy characterized by a closed feedback loop. Moreover, χ_t includes a time index to identify the day of the year. Additional variables might be considered, such as past observations of the stochastic disturbances.

Artificial Neural Networks

Using ANN to parameterize the policy, the k-th component in the control vector $\mathbf{u}_t$ (with $k = 1, \dots, n_u$) is defined as:

$$u^k_t = a_k + \sum_{i=1}^{N} b_{i,k}\psi_i(\chi_t \cdot \mathbf{c}_{i,k} + d_{i,k}) \tag{3}$$

where N is the number of neurons $\psi(\cdot)$ (i.e., hyperbolic tangent sigmoid function, which ensures universal approximation properties (Hornik et al., 1989)), $\chi_t \in \mathbb{R}^M$ the policy input vector, $a_k, b_{i,k}, d_{i,k} \in \mathbb{R}$, $\mathbf{c}_{i,k} \in \mathbb{R}^M$ the ANN parameters. The parameter vector θ is therefore defined as $\theta = [a_k, b_{i,k}, \mathbf{c}_{i,k}, d_{i,k}]$, with $i = 1, \dots, N$ and $k = 1, \dots, n_u$, and belongs to $\mathbb{R}^{n_\theta}$, where $n_\theta = n_u(N(M + 2) + 1)$.

Radial Basis Functions

In the case of RBF policy, the k-th release decision in the vector $\mathbf{u}_t$ (with $k = 1, \dots, n_u$) is defined as:

$$u^k_t = \sum_{i=1}^{N} w_{i,k}\varphi_i(\chi_t) \tag{4}$$

where N is the number of RBFs $\varphi(\cdot)$ and $w_{i,k}$ the weight of the i-th RBF. The weights are formulated such that they sum to one (i.e., $\sum_{i=1}^{N} w_{i,k} = 1$) and are non-negative (i.e., $w_{i,k} \geq 0 \quad \forall i, k$). The single RBF is defined as follows:

$$\varphi_i(\chi_t) = \exp\left[-\sum_{j=1}^{M} \frac{((\chi_t)_j - c_{j,i})^2}{b^2_{j,i}}\right] \tag{5}$$

where M is the number of input variables χ_t and $\mathbf{c}_i, \mathbf{b}_i$ are the M-dimensional center and radius vectors of the i-th RBF, respectively. The centers of the RBF must lie within the bounded input space and the radii must strictly be positive (i.e., using normalized variables, $\mathbf{c}_i \in [-1, 1]$ and $\mathbf{b}_i \in (0, 1]$). The parameter vector θ is therefore defined as $\theta = [c_{i,j}, b_{i,j}, w_{i,k}]$, with $i = 1, \dots, N$, $j = 1, \dots, M$, $k = 1, \dots, n_u$, and belongs to $\mathbb{R}^{n_\theta}$, where $n_\theta = N(2M + n_u)$.

2.2 Multi-objective evolutionary algorithms

Multi-objective evolutionary algorithms (MOEAs) are iterative search algorithms that evolve a Pareto-approximate set of solutions by mimicking the randomized mating, selection, and mutation operations that occur in nature (Coello Coello et al., 2007). These mechanisms allow MOEAs to deal with challenging multi-objective problems characterized by multi-modality, nonlinearity, and discreteness (see Nicklow et al. (2010) for an extensive review of MOEAs applications in water resources).
In this paper, we use the self-adaptive Borg MOEA (Hadka and Reed, 2013), which employs multiple search operators that are adaptively selected during the optimization, based on their demonstrated probability of generating quality solutions. In addition to adaptive operator selection, the Borg MOEA assimilates several other recent advances in the field of MOEAs, including an ε-dominance archiving with internal algorithmic operators to detect search stagnation, and randomized restarts to escape local optima. The flexibility of the Borg MOEA to adapt to challenging, diverse problems makes it particularly useful for addressing DPS problems, where the shape of the operating rule and its parameter values are problem-specific and completely unknown a priori.

3. CASE STUDY DESCRIPTION

The Red River Basin (Figure 1) is the second largest basin of Vietnam, with a total area of about 169,000 km^2, of which 48% in China, 51% in Vietnam, and the rest in Laos. Of three main tributaries, the Da River is the most important water source, contributing for 42% of the total discharge at SonTay. Since 1989, the discharge from the Da River has been regulated by the operation of the HoaBinh reservoir. With a storage capacity of 9.8 billion m^3, the HoaBinh reservoir is the largest reservoir in use in Vietnam and accounts for 15% of the national electricity production. The dam operation also contributes to flood control, especially to protect the densely populated city of Hanoi. Two operating objectives are formulated: *i*) hydropower J^{hyd}, defined in eq. (6a) as the daily average energy production (kWh/day) at the HoaBinh hydropower plant, which depends on the reservoir level h_t^{HB} and the turbined flow q_{t+1}^{Turb}; *ii*) flooding J^{flo}, defined in eq. (6b) as the daily average excess level (cm^2/day) in Hanoi with respect to the flooding threshold $\bar{h} = 950$ cm.

$$J^{hyd} = \frac{1}{H}\sum_{t=0}^{H-1} G_{t+1}(h_t^{HB}, q_{t+1}^{Turb}) \tag{6a}$$

$$J^{flo} = \frac{1}{H}\sum_{t=0}^{H-1} max(h_{t+1}^{Hanoi} - \bar{h}, 0)^2 \tag{6b}$$

The reservoir is modeled by a conceptual water-balance equation, the hydropower plant by a physically-based model, and flow routing from the reservoir to the city of Hanoi by a lumped, data-driven model. The constraints of the problem are embedded in the model (Piccardi and Soncini-Sessa, 1991), thus guaranteeing the feasibility of the designed solutions. The state variable x_t is the reservoir storage and the control u_t is the release decision. The system is affected by a stochastic disturbance vector $\mathbf{q}_{t+1}$, comprising the inflow to the reservoir q_{t+1} and the lateral flows in the Thao and Lo Rivers $q_{t+1}^{lat} = q_{t+1}^{Thao} + q_{t+1}^{Lo}$ which contribute to the flow in Hanoi. The modeling time step is 24 hours. In the adopted notation, the time subscript of a variable indicates the instant when its value is deterministically known. A data-driven feedforward neural network provides the level in Hanoi given the HoaBinh release and the tributaries' discharges. Further details about the model of the HoaBinh system can be found in Castelletti et al. (2012).

3.1 Experiment Setting

The operating policy of the HoaBinh reservoir is parameterized using ANN and RBF with different settings in terms of policy input and number of neurons/basis, as follows:

A) 3 inputs (i.e., $\sin(2\pi t/365)$, $\cos(2\pi t/365)$ and x_t) with 4 neurons/basis, $n_\theta = 21$ and 28 for ANN and RBF, respectively;
B) 4 inputs (i.e., $\sin(2\pi t/365)$, $\cos(2\pi t/365)$, x_t and q_t) with 6 neurons/basis, $n_\theta = 37$ and 54 for ANN and RBF, respectively;
C) 5 input (i.e., $\sin(2\pi t/365)$, $\cos(2\pi t/365)$, x_t, q_t and q_t^{lat}) with 8 neurons/basis, $n_\theta = 57$ and 88 for ANN and RBF, respectively;

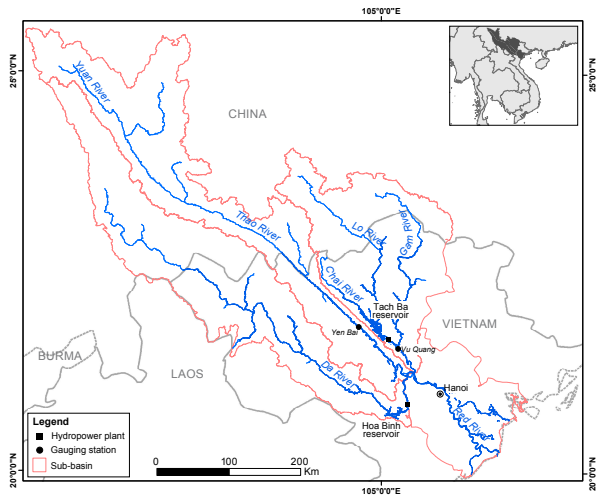

Fig. 1. The Red River basin with the HoaBinh reservoir considered in this study.

We include time t among the policy inputs by means of the terms $\sin(2\pi t/365)$ and $\cos(2\pi t/365)$, to take into account the time-dependency and cyclostationarity of the system and, consequently, of the control policy (Pianosi et al., 2011). Moreover, the past observations of the inflows q_t and q_t^{lat} allows to improve the control ability of the policy enlarging the information on the system condition. The number of neurons/basis increases with the number of policy inputs and was fixed via trial-and-error. The dimensionality of the problem increases moving from setting A to settings B and C, where DP family methods cannot be applied as the state vector comprises the time index, the reservoir storage, and the additional states required to model the inflows q_t and q_t^{lat}.
Since the Borg MOEA has been demonstrated to be relatively insensitive to the choice of parameters, we use the default algorithm parameterization suggested by Hadka and Reed (2013). Each optimization was run for 200,000 function evaluations. To improve solution diversity and avoid dependence on randomness, the solution set from each formulation is the result of 10 random optimization trials. In total, each optimization comprises 2 million simulations over the horizon 1962-1969 and requires approximately 5 hours on a 2 processors Intel Xeon E5-2660 2.20 GHz with 96 GB Ram. The performance of the resulting policies is then computed over the validation horizon 1995-2004, with the final set of Pareto-optimal policies for each input settings defined as the set of non-dominated solutions from the results of all the optimization trials.

4. APPLICATION RESULTS

Figure 2 reports the approximated Pareto front obtained with the two universal approximators (i.e., ANN and RBF) for the three input settings simulated over the validation horizon (1995-2004). The arrows show the direction of preference for each objective, with the ideal solution in the top-left corner of each panel. The approximated Pareto front obtained with RBF policies outperforms the one obtained with ANN policies over all the input sets. In general, RBF policies allow a better exploration of the tradeoff between J^{flo} and J^{hyd}, with the RBF compromise solutions Pareto-dominating the corresponding ANN

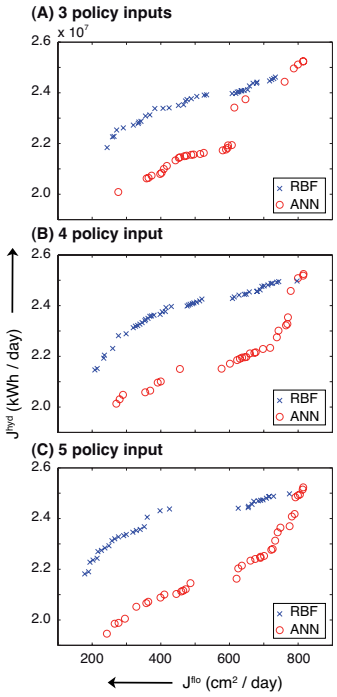

Fig. 2. Pareto front approximations with different settings in terms of control policy input.

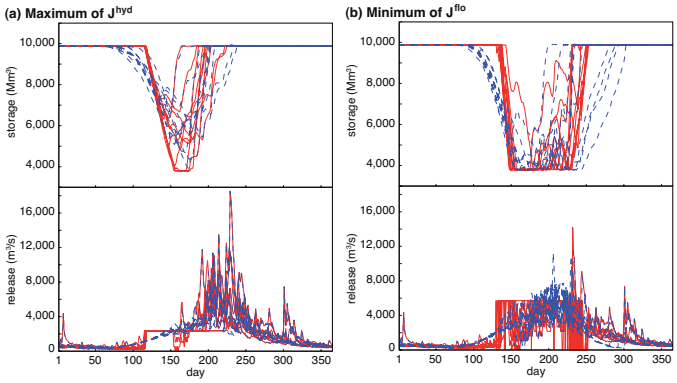

Fig. 3. Yearly patterns of the HoaBinh storage and release over the validation horizon (1995-2004) under the best ANN (red) and RBF (blue) policy for hydropower production (panel (a)) and flood control (panel (b)) for input setting C.

solutions. At the two extremes of the Pareto front, the performance of ANN and RBF is similar. ANN policies indeed attain the same maximum of hydropower production of RBF policies (see the top-right part of each panel), with a limited improvement in experiment setting A. This is confirmed in Figure 3a, which shows that the yearly patterns of the reservoir storage and release under the ANN and RBF policies attaining the maximum of J^{hyd} are rather similar (input setting C). Conversely, although both ANN and RBF attain approximately the same minimum value of the flooding objective of J^{flo} (see the bottom-left part of each panel in Figure 2), RBF policies are able to obtain this performance at higher levels of hydropower production. The reason can be inferred from Figure 3b, which shows that the best RBF policy for flood control (blue) stores more water in the summer period than the

Table 1. Metrics values of the approximated Pareto front shown in Figure 2.

		Gen. Distance	ε-indicator	Hypervolume
A)	ANN	0.0824	0.476	0.199
	RBF	0.0207	0.183	0.491
B)	ANN	0.0952	0.533	0.0921
	RBF	0.0113	0.0998	0.579
C)	ANN	0.0787	0.531	0.110
	RBF	0.000713	0.0797	0.658

best ANN policy (red) and therefore can continuously guarantee some hydropower release, while the ANN policy often generates very low releases in this period.

The quality of the two approximate Pareto fronts in Figure 2 is further analyzed by means of three formal metrics, namely *generational distance*, *additive ε-indicator*, and *hypervolume*, which respectively account for convergence, consistency, and diversity (Knowles and Corne, 2002; Zitzler et al., 2003). In principle, these metrics should be computed with respect to the optimal Pareto front. In practice, since the computation of the optimal Pareto front via DP might be impracticable due to the dimensionality of the state and objective function vectors, its best approximation (also called reference set) is considered. This latter is constructed by selecting all the solutions non Pareto dominated obtained in any optimization trials. Table 1 reports the three metrics evaluated for the different approximated Pareto front shown in Figure 2. The values of the metrics confirm the superiority of the RBF parametrization of the control policies over the ANN one. Hypervolume values of RBF approximations are three to six times greater than the ones obtained by ANN. Similarly, the ε-indicator values for RBF are significantly lower than the ones for ANN, due to multiple gaps in some tradeoff regions. Finally, also the generational distance confirms that RBF solutions outperform the ANN ones.

The superiority of RBF over ANN policies may be explained by the different definition of the parameter space in which the solutions are searched for. In the case of RBF, the parameter space is the Cartesian product of the subsets $[-1, 1]$ for each center $c_{j,i}$ and $(0, 1]$ for each radius $b_{j,i}$ and weight $w_{i,k}$. In the case of ANN, instead, parameters have no direct relationship with the policy inputs. In this work, the domain $-10000 < a_k, b_{i,k}, \mathbf{c}_{i,k}, d_{i,k} < 10000$ has been used as in Castelletti et al. (2013b). Although this large domain should guarantee flexibility to the ANN structure and prevents that any Pareto-optimal solution be excluded a priori, it makes the search more difficult and slow down convergence. It is worth noting that the difference in the number of parameters (i.e., 21 for ANN and 28 for RBF with setting A, which increases to 57 for ANN and 88 for RBF with setting C) seems instead to play a minor role.

5. CONCLUSIONS

The paper presents a comparative analysis of two universal approximators, namely Artificial Neural Networks and Radial Basis Functions, in multi-objective Direct Policy Search problems. The regulation of the multi-purpose HoaBinh water reservoir in Vietnam is used as a case study. The combination of DPS with the Borg MOEA shows the potential to overcome the limitation of dynamic programming family methods. The proposed method successfully solves high-dimensional state and control space

problems and finds an approximation of the entire Pareto front, and the associated control policies, in a single optimization run. The comparison of ANN and RBF control policy parameterizations suggests the superiority of RBF, which successfully explores the full tradeoff space between the two conflicting objectives. This result can be probably explained by the smaller parameters space of RBF with respect to the one of ANN. Although accurate tuning and preconditioning of ANN policies improve the performance (Castelletti et al., 2013b), they require a priori information about the shape of the optimal policy. RBF policies attain good results without any preconditioning, thus representing a potentially effective, case-independent option.

Future research will focus on estimating the sensitivity of each parameterization to the underlying architecture (e.g., comparison of ANN and RBF with equal number of parameters and varying number and type of neurons/basis). Moreover, we will extend the comparative analysis by including other approximators, such as fuzzy systems or support vector machine, and by testing the scalability of each control policy approximation also with respect to the dimension of the output (control) vector.

REFERENCES

Bellman, R. (1957). *Dynamic programming*. Princeton University Press, Princeton.

Busoniu, L., Babuska, R., De Schutter, B., and Ernst, D. (2010). *Reinforcement learning and dynamic programming using function approximators*. CRC Press.

Castelletti, A., Pianosi, F., Quach, X., and Soncini-Sessa, R. (2012). Assessing water reservoirs management and development in Northern Vietnam. *Hydrology and Earth System Sciences*, 16(1), 189–199. doi:10.5194/hess-16-189-2012.

Castelletti, A., Pianosi, F., and Restelli, M. (2013b). A multiobjective reinforcement learning approach to water resources systems operation: Pareto frontier approximation in a single run. *Water Resources Research*, 49. doi: 10.1002/wrcr.20295.

Castelletti, A., Yajima, H., Giuliani, M., Soncini-Sessa, R., and Weber, E. (2013a). Planning the Optimal Operation of a Multi–Outlet Water Reservoir with Water Quality and Quantity Targets. *Journal of Water Resources Planning and Management*. doi: 10.1061/(ASCE)WR.1943-5452.0000348.

Coello Coello, C., Lamont, G., and Veldhuizen, D.V. (2007). *Evolutionary Algorithms for Solving Multi-Objective Problems (Genetic Algorithms and Evolutionary Computation)*. Springer, New York, 2 edition.

Deisenroth, M., Neumann, G., and Peters, J. (2011). A Survey on Policy Search for Robotics. In *Foundations and Trends in Robotics*, volume 2, 1–142.

Hadka, D. and Reed, P. (2012). Diagnostic assessment of search controls and failure modes in many–objective evolutionary optimization. *Evolutionary Computation*, 20(3), 423–452.

Hadka, D. and Reed, P. (2013). Borg: An Auto–Adaptive Many–Objective Evolutionary Computing Framework. *Evolutionary Computation*, 21(2), 231–259.

Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5), 359–366.

Knowles, J. and Corne, D. (2002). On metrics for comparing non–dominated sets. In *Proceedings of the 2002 World Congress on Computational Intelligence (WCCI)*, 711–716. IEEE Computer Society.

Koutsoyiannis, D. and Economou, A. (2003). Evaluation of the parameterization-simulation-optimization approach for the control of reservoir systems. *Water Resources Research*, 39(6), 1170–1187.

Kurková, V. and Sanguineti, M. (2001). Bounds on rates of variable-basis and neural-network approximation. *IEEE Transactions on Information Theory*, 47(6), 2659–2665.

Liao, Y., Fang, S., and Nuttle, H. (2003). Relaxed conditions for radial-basis function networks to be universal approximators. *Neural Networks*, 16(7), 1019–1028.

Lund, J. and Guzman, J. (1999). Derived operating rules for reservoirs in series or in parallel. *Journal of Water Resources Planning and Management*, 125(3), 143–153.

Nicklow, J., Reed, P., Savic, D., Dessalegne, T., Harrell, L., Chan-Hilton, A., Karamouz, M., Minsker, B., Ostfeld, A., Singh, A., and Zechman, E. (2010). State of the Art for Genetic Algorithms and Beyond in Water Resources Planning and Management. *Journal of Water Resources Planning and Management*, 136(4), 412–432. doi:10.1061/(ASCE)WR.1943-5452.0000053.

Peters, J. and Schaal, S. (2008). Reinforcement learning of motor skills with policy gradients. *Neural networks*, 21(4), 682–697.

Pianosi, F., Quach, X., and Soncini-Sessa, R. (2011). Artificial Neural Networks and Multi Objective Genetic Algorithms for water resources management: an application to the Hoabinh reservoir in Vietnam. In *Proceedings of the 18th IFAC World Congress*. Milan, Italy.

Piccardi, C. and Soncini-Sessa, R. (1991). Stochastic dynamic programming for reservoir optimal control: Dense discretization and inflow correlation assumption made possible by parallel computing. *Water Resources Research*, 27(5), 729–741.

Powell, W. (2007). *Approximate Dynamic Programming: Solving the curses of dimensionality*. Wiley, NJ.

Rosenstein, M. and Barto, A. (2001). Robot weightlifting by direct policy search. In *International Joint Conference on Artificial Intelligence*, volume 17, 839–846. Citeseer.

Tikk, D., Kóczy, L., and Gedeon, T. (2003). A survey on universal approximation and its limits in soft computing techniques. *International Journal of Approximate Reasoning*, 33(2), 185–202.

Tsitsiklis, J. and Van Roy, B. (1996). Feature-Based Methods for Large Scale Dynamic Programming. *Machine Learning*, 22, 59–94.

White, D. (1982). Multi-objective infinite-horizon discounted Markov decision processes. *Journal of Mathematical Analysis and Optimization*, 89(2), 639–647.

Zitzler, E., Thiele, L., Laumanns, M., Fonseca, C., and da Fonseca, V. (2003). Performance assessment of multiobjective optimizers: an analysis and review. *IEEE Transactions on Evolutionary Computation*, 7(2), 117–132.