GSC Advanced Research and Reviews
eISSN: 2582-4597 CODEN (USA): GARRC2
Cross Ref DOI: 10.30574/gscarr
Journal homepage: https://gscarr.gsconlinepress.com/

(RESEARCH ARTICLE)

INFLUENCE OF TANK SIZE ON THE TECHNO-ECONOMIC SIZING OF A PHOTOVOLTAIC WATER PUMPING SYSTEM: A HYBRID ARTIFICIAL INTELLIGENCE APPROACH APPLIED TO THE SAHEL (SAKAL, SENEGAL)

Souleymane SOUMAH [1, 2, 3, *], Mouhamadou Falilou NDIAYE [1], Mohamed Ansoumane CAMARA [2, 3], Diatta SENE [1], Kadiatou Aissatou BARRY [2] and Vincent SAMBOU [1]

[1] Water, Energy, Environment and Industrial Processes Laboratory (LE3PI), Cheikh Anta Diop University, Dakar, Senegal.
[2] Laboratory for Teaching and Research in Energy, Automation and Telecommunications (LENAT), UGANC, Conakry Guinea.
[3] Department of Electrical Engineering of the Polytechnic Institute, University Gamal Abdel Nasser de Conakry (UGANC), Conakry, Guinea.
* Corresponding Author

ORCID Details
Souleymane SOUMAH : https://orcid.org/0009-0002-1280-5570
Mouhamadou Falilou NDIAYE : https://orcid.org/0000-0002-9421-7598
Mohamed Ansoumane CAMARA: https://orcid.org/0009-0009-0922-3415
Diatta SENE: https://orcid.org/0009-0000-1725-3870
Kadiatou Aissatou BARRY: https://orcid.org/0009-0005-0204-9021
Vincent SAMBOU: https://orcid.org/0000-0002-1042-8348

GSC Advanced Research and Reviews, 2026, 29(01), 343–360

Publication history: Received on 10 September 2026; revised on 24 October 2026; accepted on 26 October 2026

Article DOI: https://doi.org/10.30574/gscarr.2026.29.1.0288

Copyright © 2026 Author(s) retain the copyright of this article. This article is published under the terms of the Creative Commons Attribution License 4.0

ABSTRACT

In rural Sahelian areas, access to water relies heavily on battery-less photovoltaic (PV) pumping, in which the tank acts as the storage element. Its volume, however, is most often set by autonomy rules of one to three days, which are unrelated to the installed power and to the variability of solar radiation. This paper quantifies the influence of this volume on techno-economic sizing using a hybrid chain that combines a Gradient Boosting power model, a climate-sequence generator based on typical day types and a Markov chain, a surrogate model (neural networks and Gradient Boosting), and a genetic algorithm with an uncertainty margin, with verification by simulation. The case study concerns Sakal (Louga, Senegal, Sahelian climate) and uses three days measured in 2020 at a 15-min time step. For a demand of 400 m^3/day, a head of 40 m, and a tolerated deficit of 5 %, the optimum is a 225 m^3 tank (0.56 day of autonomy) with 23 kWp, at 0.108 USD/m^3. Autonomy rules of one, two, and three days raise the cost of water by 34, 112, and 183 %. Without a tank, the deficit exceeds 52 % regardless of the PV power. The optimal autonomy remains between 0.50 and 0.59 day across 17 sensitivity scenarios. These results depend on the assumed unit costs and on a limited set of measurements; their validation on multi-year local data is necessary.

Keywords: Photovoltaic Water Pumping; Tank Sizing; Levelized Cost of Water (LCOW); Water Supply Deficit Probability; Hybrid Artificial Intelligence; Genetic Algorithm; Sahel.

1 INTRODUCTION

Access to safely managed drinking water remains highly unequal: according to the 2025 report of the WHO/UNICEF Joint Monitoring Programme (JMP), 2.1 billion people still lacked it in 2024, although global coverage rose from 68 % to 74 % between 2015 and 2024. The largest disparities affect low-income countries, fragile contexts, and rural areas [1]. In rural localities far from the electricity grid, notably in the Sahel, photovoltaic (PV) pumping has become a major technical and economic solution, because solar radiation is abundant there and water demand is largely diurnal [2] ,[3].

In the "sun-following" configuration, the tank acts as a hydraulic battery: it replaces electrochemical storage, whose use significantly increases the system cost [4]. Its effect on reliability is considerable. For a village in western Egypt, the annual probability of supply failure falls from 0.38 without a tank to 0.005 with a 71 m^3 tank, for a small increase in life-cycle cost [5]. A feasibility study concludes that increasing the maximum tank capacity is, technically and economically, preferable to adding another energy source or additional storage [6]. A method for estimating the loss-of-load probability (LLP) of PV pumping systems was proposed as early as 2004, for a tank with two days of autonomy; it makes it possible to draw LLP maps from observed or generated meteorological series [7].

In practice, however, the tank size is often set by empirical rules expressed in days of autonomy: a common rule is to provide three days without solar energy [8]. These rules take into account neither the installed power nor the variability of solar radiation. An integrated sizing method made it possible, in one case study, to reduce the required PV power by 20 % and the storage capacity by 50 % compared with the traditional method [9], which suggests that the usual rules may oversize the storage. Optimal sizing has been addressed through the loss-of-power-supply probability (LPSP) criterion and nature-inspired algorithms [10],[11], and then through more recent metaheuristics. Mazloumi et al. jointly size the generator and the tank by minimizing the life-cycle cost under a water-interruption constraint [4], while Abulkheir et al. combine cost, LLP, and PV curtailment using particle swarm optimization [5]. Mahmoud et al. emphasize that inaccurate sizing leads to oversizing and financial waste, and propose an online pre-sizing tool [12]. Liang and Shorter show, for smallholder farms in sub-Saharan Africa, that the pump efficiency determines the size and cost of the PV array [13].

These studies nevertheless leave three limitations. First, the effect of the tank volume is often evaluated on daily profiles: Mazloumi et al., for example, use irradiance and temperature profiles for a single day [4], which does not allow the inter-day carry-over during sequences of low-sunshine days to be represented. Markov-chain generators of day sequences [14] meet this need but are little used in the joint optimization of the generator and the tank. Second, uncertainty is mostly handled through one-at-a-time sensitivity analyses; Mazloumi et al. themselves suggest evaluating the uncertainty of PV production and water consumption as a future direction [4]. Third, the computational cost of multi-year, fine-time-step evaluations repeated inside the optimization loop is rarely addressed.

Artificial intelligence (AI) provides partial answers to these limitations. For power generation, a 2024 comparison of short-term irradiance forecasting ranks Gradient Boosting regression first among six machine-learning models [15], and another evaluation compares six algorithms for predicting solar power [16]. For optimization, genetic algorithms [17] and their multi-objective versions [18] are common, and surrogate models reduce the number of costly simulations [19], [20],[21],[22]. In the design of pumping stations, Faúndez-Lizama et al. couple an evolutionary algorithm with a machine-learning classifier and reduce the number of objective-function evaluations from 50,000 to fewer than 25,000; they acknowledge, however, that the model must be retrained for each network [23]. Surrogates can rely on Gradient Boosting [24] or on neural networks [25]. To our knowledge, the combination of an AI-based power model, a surrogate covering several design parameters, and a genetic algorithm that accounts for the surrogate's uncertainty has not been applied to the joint sizing of the generator and the tank.

The objective of this paper is to establish the influence of the tank volume on the techno-economic sizing of a battery-less PV pumping system, and to propose a reproducible decision-support methodology. The contributions are as follows: a hybrid AI chain with three modules (a Gradient Boosting power model, a surrogate combining neural networks and Gradient Boosting that covers the main design parameters, and a constrained genetic algorithm with an uncertainty margin, complemented by a cost–reliability Pareto front); a stochastic representation of the solar resource through typical day types and a Markov chain, which makes inter-day storage visible; and a sensitivity analysis that measures the robustness of the sizing to technical and economic assumptions. The methodology is illustrated by a case study in a Sahelian climate, at the site of Sakal, a rural locality in northwestern Senegal (Louga region), whose data and results are presented in a dedicated section. The paper is organized as follows: Section 2 presents the system modeling, Section 3 the hybrid methodology, Section 4 the case study, Section 5 the results and their discussion, and Section 6 concludes.

2 SYSTEM MODELING

2.1 System Architecture

The system studied is a battery-less photovoltaic (PV) pumping system (Fig. 1). A PV generator supplies, through a power-conditioning unit (a maximum power point tracking converter and an inverter), a submersible pump installed in a borehole. The water is delivered to an elevated tank that feeds the distribution network by gravity. This tank-based architecture is the most widely used for domestic PV pumping, the battery-bank architecture being rarely used [26], [27].

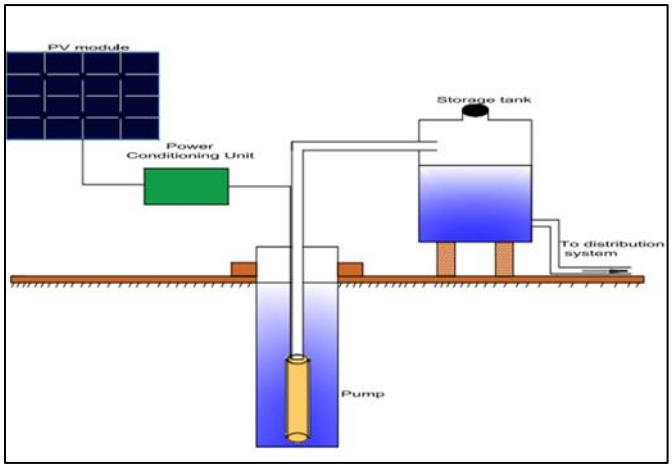


Figure 1: Schematic of the battery-less photovoltaic pumping system: PV generator, power-conditioning unit, submersible pump, storage tank, and outlet to the distribution network.

2.2 Modeling Assumptions

The following assumptions are adopted in the remainder of the study:

- **(H1)** ***Maximum power point tracking.*** The generator operates at its maximum power point. The efficiencies of the converter, the motor, and the pump are lumped into an overall efficiency η (Section 2.4).
- **(H2)** ***Constant total head.*** The total dynamic head H (dynamic level, delivery height, and head losses) is assumed constant. This simplification is supported by a study of a borehole in Burkina Faso, which shows that the drawdown caused by pumping has a negligible effect, whereas the variation of the static level may require up to +100 % of peak power [28]. The uncertainty on H is therefore handled through a sensitivity analysis (Section 5).
- **(H3)** ***Deterministic demand.*** The daily demand D is constant and distributed according to a fixed intra-day profile (Section 2.5).
- **(H4)** ***Fine time step.*** The computation time step is $\Delta t = 15$ min. The temporal resolution of the consumption profile has a strong impact on the modeling and on the optimal sizing [29], whereas hourly satellite data can replace a local sensor without noticeably changing the accuracy or the optimum [30]. The 15-min step is dictated here by that of the available measurements.
- **(H5)** ***No degradation.*** Variable head losses, module degradation, and pump aging are neglected. The effect of temperature on production is embedded in the measurements.

2.3 Photovoltaic Generator

To first order, the power of a module increases with irradiance and decreases with cell temperature, according to well-documented algebraic forms [31]. Rather than modeling the physics of each cell, we start from the measured power $P_{\mathrm{ref}}(t)$ of a reference generator of peak power $P_{\mathrm{c,ref}}$, and scale it by a dimensionless factor s (Equations 1 and 2):

$$P_{\mathrm{pv}}(t) = s \cdot P_{\mathrm{ref}}(t) \qquad (1)$$

$$P_{\mathrm{c}} = s \cdot P_{\mathrm{c,ref}} \qquad (2)$$

This scaling assumes that the sized generator has the same technology, orientation, and exposure conditions as the reference generator. The relationship between power and irradiance is learned by an artificial intelligence model (Section 3), which makes it possible to generate the power of unmeasured days.

2.4 Conversion Chain and Pump

The nominal power of the motor-pump unit is taken equal to that of the generator, $P_{\mathrm{n}} = P_{\mathrm{c}}$.

The pump load ratio is defined by Equation (3):

$$x(t) = \min\left[\frac{P_{\mathrm{pv}}(t)}{P_{\mathrm{n}}}, 1\right] \qquad (3)$$

The overall efficiency of the chain (converter, motor, pump) decreases at part load. It is represented by the empirical parametric form in Equation (4):

$$\eta(x) = \begin{cases} \eta_{\mathrm{nom}} \cdot [1 - \kappa(1-x)^2], & \text{if } x \geq x_{\min} \\ 0, & \text{otherwise} \end{cases} \qquad (4)$$

where η_{nom} is the efficiency at nominal load, κ the part-load loss coefficient, and $x_{\min}$ the minimum start-up load ratio.

This form is a modeling choice and not the result of a reference. It aims to reproduce qualitatively the drop in efficiency as irradiance weakens; its parameters will be specified or calibrated in the case study. The volume pumped during a time step Δt is written as

$$\Delta V_p(t) = \frac{\eta(x) \cdot \min[P_{\mathrm{pv}}(t),\, P_n] \cdot \Delta t}{\rho \cdot g \cdot H} \qquad (5)$$

where ρ is the density of water and g the acceleration due to gravity.

This relation reflects the fact that the flow rate of a PV pumping system is a function of the available power: an integrated sizing method, moreover, relies on such a flow–power function [9]. An experimentally validated model of a real system also simulates the pumped flow rate and the tank level from the irradiance, the ambient temperature, and the users' withdrawal profile [32]. Recent work on controlling the operating point of the motor shows that it is possible to improve the quantity of water pumped per day and the system efficiency [33], which justifies treating η as a design parameter to be characterized.

2.5 Water Demand

The daily demand D is distributed according to an intra-day profile φ_k satisfying $\sum_k \varphi_k = 1$ over the 96 time steps of the day. Measurements on a rural household show that hourly consumption has clearly marked peaks in the morning and in the evening, with a very similar distribution on every day of the week [34]. This profile is modeled as a sum of Gaussian peaks (morning, midday, evening) on a constant background, and then normalized:

$$\varphi_k = \frac{f(t_k)}{\sum_j f(t_j)} \qquad (6a)$$

$$f(t) = \sum_m a_m \exp\left[-\frac{(t-\mu_m)^2}{2\sigma_m^2}\right] + b \qquad (6b)$$

The number and shape of the peaks are **a modeling choice**; the parameters a_m, μ_m, σ_m, and b will be calibrated on the site in the case study. The demand at time step k is $D\,\varphi_k$.

2.6 Tank Balance and Performance Indicators

For a tank of volume V, the stored volume V_k at time step k evolves according to

$$V_{k+1} = V_k + \Delta V_{\mathrm{in},k} - \Delta V_{\mathrm{out},k} \qquad (7)$$

$$\Delta V_{\mathrm{in},k} = \min\left(\Delta V_{\mathrm{p},k}, V - V_k\right) \qquad (8a)$$

$$\Delta V_{\mathrm{out},k} = \min\left(V_k + \Delta V_{\mathrm{in},k}, D\,\varphi_k\right) \qquad (8b)$$

When the tank is full, the pump is curtailed and the corresponding energy is lost. When it is empty, the unmet demand constitutes a deficit. The initial level is set to a fraction v_0 of V at the beginning of each simulated year. Two indicators are defined over the whole simulated period:

$$\mathrm{WSPP} = \frac{\sum_k (D\,\varphi_k - \Delta V_{\mathrm{out},k})}{\sum_k D\,\varphi_k} \quad (9)$$

$$c = \frac{\sum_k (\Delta V_{\mathrm{p},k} - \Delta V_{\mathrm{in},k})}{\sum_k \Delta V_{\mathrm{p},k}} \quad (10)$$

The water supply deficit rate, WSPP (*Water Supply Probability of Deficiency*), is the share of unmet demand. It is akin to the loss-of-load probability (LLP), introduced for PV pumping systems as early as 2004 [7], and to the water-interruption probability used as a constraint in recent work [4], but here it is volumetric and not based on counting deficit time steps. The unused-energy rate c measures the share of the pumpable volume lost through curtailment; it is comparable to the curtailed PV power indicator used in other studies [5].

2.7 Economic Model

The cost of water is expressed by the levelized cost of water (LCOW), an approach derived from the life-cycle cost analysis used in PV pumping studies, including to express a cost per cubic meter of water [4],[27],[35]. The initial investment comprises the generator, the motor-pump unit, the tank, and a fixed cost C_{fix} (borehole and civil works excluding the tank):

$$C_0 = c_{\mathrm{pv}} P_{\mathrm{c}} + c_{\mathrm{p}} P_{\mathrm{n}} + a\,V^b + C_{\mathrm{fix}} \quad (11)$$

where c_{pv} and c_{p} are the unit costs of the generator and of the motor-pump unit. The tank cost follows a power law with exponent $b < 1$, which reflects economies of scale: the storage cost per liter decreases as the size increases [36]. The parameters a and b will be calibrated on local costs. The annual operating costs are fractions m_{pv}, m_{p}, and m_{r} of the initial costs, and the motor-pump unit is replaced in year t_{r} for a fraction ρ_{r} of its initial cost:

$$C_{\mathrm{om}} = m_{\mathrm{pv}} c_{\mathrm{pv}} P_{\mathrm{c}} + m_{\mathrm{p}} c_{\mathrm{p}} P_{\mathrm{n}} + m_{\mathrm{r}} a\,V^b \quad (12)$$

With a discount rate r and a lifetime N, the cumulative discount factor A and the annual delivered volume W are

$$A = \frac{1 - (1+r)^{-N}}{r} \quad ,$$

$$W = 365 \cdot D \cdot (1 - \mathrm{WSPP})$$

- where:
- D: daily water demand,
- $WSPP$: system deficit rate (Water Supply Probability of Deficiency).

This relation expresses that the annual delivered volume corresponds to the total demand over 365 days, corrected by the deficit rate.

The levelized cost of water is written as Equation (13):

$$\mathrm{LCOW} = \frac{C_0 + A\,C_{\mathrm{om}} + \rho_{\mathrm{r}}\,c_{\mathrm{p}} P_{\mathrm{n}} (1+r)^{-t_{\mathrm{r}}}}{A\,W} \quad (13)$$

2.8 Formulation of the Optimization Problem

Sizing consists in choosing the volume V and the power factor s that minimize the levelized cost of water under a reliability constraint:

$$\min_{V,s} \mathrm{LCOW}(V,s) \quad (14a)$$

$$\mathrm{WSPP}(V,s) \le \varepsilon, \qquad V_{\min} \le V \le V_{\max}, \qquad s_{\min} \le s \le s_{\max} \quad (14b)$$

The function $\mathrm{WSPP}(V,s)$ is accessible only by simulating the balance (7)–(8) over several years at a 15-min time step, which motivates the hybrid methodology presented in Section 3. Table 1 summarizes the notation.

Table 1: Notation of Section 2

Symbol	Definition	Unit
s	Generator scaling factor (installed peak power $= s\,P_{\mathrm{c,ref}}$)	—
P_{c}, P_{n}	Generator peak power; pump nominal power	kW
x, $x_{\min}$	Pump load ratio; minimum start-up ratio	-
η, η_{nom}, κ	Overall efficiency; nominal efficiency; part-load loss coefficient	-
H	Total dynamic head	m
D, φ_k	Daily demand; fraction of the demand at time step k	m^3/j ; -
V, V_k	Tank volume; stored volume at time step k	m^3
WSPP, c	Water supply deficit rate; unused-energy rate	-
C_0, C_{om}	Initial cost; annual operating cost	USD
r, N, A	Discount rate; lifetime; cumulative discount factor	years
LCOW	Levelized cost of water	USD/m^3

3 HYBRID METHODOLOGY

3.1 Overview

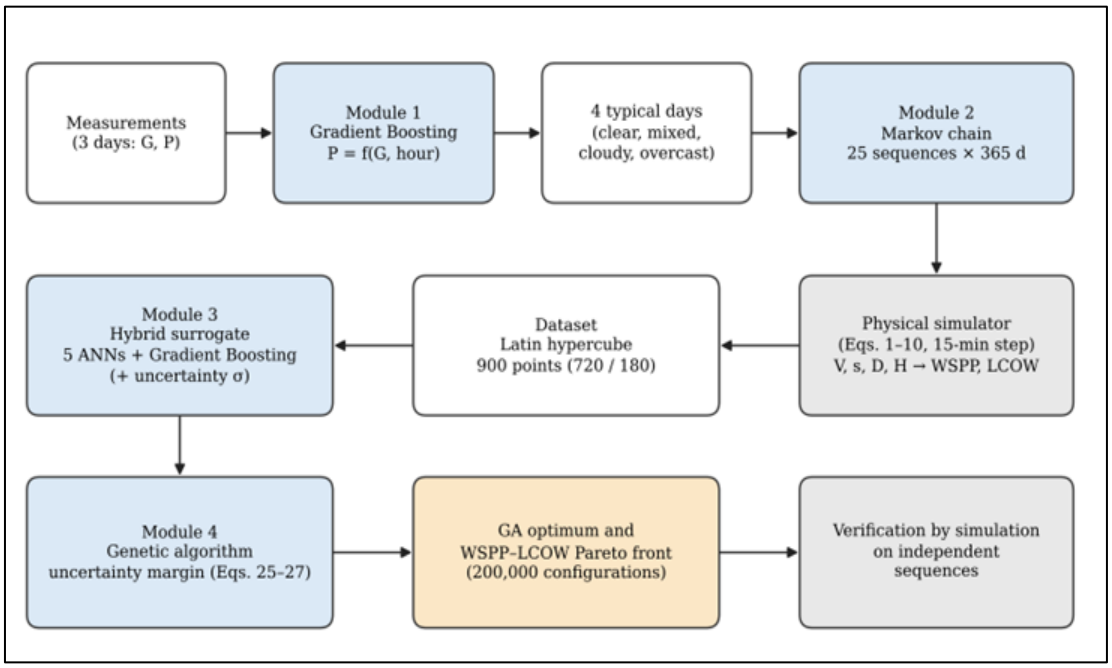


Figure 2: Hybrid methodological chain. The artificial intelligence modules are in blue; the simulator and the verification are in gray.

Problem (14) is difficult to handle directly, because the reliability WSPP(V, s) is accessible only by simulation. One evaluation runs through 25 simulated years at a 15-min time step, i.e., 25 × 365 × 96 = 876,000 computation steps. Exploring the space (V, s) with an evolutionary algorithm that calls the simulator directly is possible, as in multi-objective sizing studies of PV pumping systems [24] , but the cost becomes prohibitive as soon as demand, head, and climate uncertainty are added to the study variables. Surrogate approaches, which replace the simulation with a learning model trained on simulations, are emerging for the sizing of renewable energy systems [37],[38], but, to our knowledge, not for the tank volume–peak power pair of a battery-less PV pumping system.

The proposed methodology articulates four artificial intelligence modules around the physical simulator of Section 2 (Fig. 2): (i) a *Gradient Boosting* power model that learns the irradiance–power relationship from the measurements; (ii) a climate generator based on typical day types and a Markov chain, which produces long sequences from a few measured

days; (iii) a hybrid surrogate, composed of an ensemble of neural networks and a *Gradient Boosting* model, which predicts WSPP at low cost; (iv) a genetic algorithm that optimizes the sizing with an uncertainty margin. The simulator remains the arbiter: any solution proposed by the surrogate is verified by simulation on sequences independent of those used for the optimization. Table 2 groups the parameters of the chain.

3.2 Module 1: Gradient Boosting Power Model

The available measurements cover a limited number of days. To produce the power of unmeasured days, for example a cloudier day, a model that relates the power P_k to the irradiance G_k is needed. A linear regression $P = \beta G$ cannot represent the nonlinearities (inverter start-up threshold, effect of cell temperature). We adopt *Gradient Boosting* [24], an ensemble method based on regression trees that has proven its worth for PV power forecasting, including in combination with a prior clustering of the data by weather type [39] or with a signal decomposition [40]. The input vector at time step k is

$$z_k = \left(G_k, h_k, \sin\left(\frac{2\pi h_k}{24}\right), \cos\left(\frac{2\pi h_k}{24}\right)\right) \qquad (15)$$

where h_k is the hour of the day. The sine and cosine terms give the hour a periodic encoding and capture the diurnal effects (temperature, angle of incidence) that irradiance alone does not carry. The model is built additively from M shallow trees T_m:

$$F_m(z) = F_{m-1}(z) + \nu\, T_m(z), \qquad m = 1, \ldots, M \qquad (16)$$

$$T_m \qquad \text{fit} \qquad r_{i,m} = P_i - F_{m-1}(z_i) \qquad (17)$$

In Equation (16), F_0 is the mean of the training powers and ν the learning rate; in Equation (17), $r_{i,m}$ is the residual of the previous iteration, i.e., the negative gradient of the squared loss. Each tree is fitted on a random subsample of the data, which reduces overfitting. The implementation is that of *scikit-learn* [41]. The predicted power is $P_{\text{pred}} = F_M(z)$, bounded below by zero.

Since the measurements are few, the quality of the model is assessed by leave-one-day-out cross-validation: the model is trained on $n - 1$ days and tested on the excluded day, with the root-mean-square error and the coefficient of determination R^2 as criteria. Linear regression serves as the baseline. A tree cannot extrapolate beyond the observed irradiance range; this choice is consistent with the scaling (1), which preserves the shape of the measured power.

3.3 Module 2: Typical Days and Markov Chain

A long climate series is not available. We replace it with a two-level stochastic generator, in the spirit of sky-state models described by a Markov chain, a recent application of which to the sizing of stand-alone solar systems is proposed in [42]. The first level is a set of $n_t = 4$ typical day types (clear, mixed, cloudy, overcast), obtained from the measurements and from the model of Module 1. A classification of days into types combined with a Markov chain has also been used to build typical solar years [43]. The second level describes the succession of types from one day to the next. Denoting by c_d the type of day d, the transition matrix is

$$T_{ij} = \Pr(c_{d+1} = j \mid c_d = i) = \rho\, \delta_{ij} + (1 - \rho)\, \pi_j \qquad (18)$$

where $\pi = (\pi_1, \ldots, \pi_{n_t})$ is the frequency of each type, δ_{ij} the Kronecker delta, and $\rho \in [0, 1)$ the persistence parameter. This form guarantees that π is the stationary distribution of the chain:

$$\pi\, T = \rho\, \pi + (1 - \rho)\, \pi = \pi \qquad (19)$$

The persistence ρ controls the duration of sequences of similar days, and therefore the length of the unfavorable periods that the tank must absorb. At $\rho = 0$, the types are drawn independently; as ρ increases, overcast episodes become longer. From T, we simulate $N_R = 25$ independent 365-day replications. The power of day d at time step k is that of the drawn type:

$$P_{\text{ref}}(d, k) = P_k^{(c_d)} \qquad (20)$$

A second set of sequences, with a different seed, is used exclusively for validation. The frequencies π and the persistence ρ are modeling assumptions; their influence will be quantified by the sensitivity analysis (Section 3.7). In the case study, they should ideally be estimated from a local climate series.

3.4 Simulation Dataset

The surrogate is trained on simulations sampled by Latin hypercube [44], which spreads the points uniformly in a space of any dimension and reduces the number of simulations required compared with random sampling; the combination of this type of sampling and a neural network is a common practice in surrogate-based optimization [45]. The parameter vector is

$$\theta = (\ln V,\ s,\ D,\ H) \in \Theta \quad (21)$$

where Θ is a bounded box (Table 2). The volume is sampled on a logarithmic scale, because its effect on reliability varies in order of magnitude rather than in absolute value. For each of the $N = 900$ points, the simulator provides $\mathrm{WSPP}(\theta)$ over the 25 optimization sequences. The dataset is split into 720 training points and 180 test points. The demand D and the head H are included in the inputs so that the same surrogate can serve the sensitivity analysis.

3.5 Module 3: Hybrid Surrogate

The deficit WSPP is bounded between 0 and 1 and concentrates near 0. To avoid negative predictions and to better resolve the low-deficit zone, the surrogate learns the square root of the target, $y = \sqrt{\mathrm{WSPP}}$. It combines two families of models. The first is an ensemble of $K = 5$ neural networks $m_j(\theta)$, with two hidden layers of 48 neurons and hyperbolic tangent activation, initialized with different seeds. The second is a *Gradient Boosting* model $g(\theta)$, of the same type as that of Module 1. The hybrid prediction is their mean, and the ensemble of networks provides a measure of dispersion:

$$\mu(\theta) = \frac{1}{2}\,\mu_{\mathrm{ANN}}(\theta) + \frac{1}{2}\,g(\theta), \qquad \mu_{\mathrm{ANN}} = \frac{1}{K}\sum_{j=1}^{K} m_j(\theta) \qquad (22)$$

$$\sigma(\theta) = \sqrt{\frac{1}{K}\sum_{j=1}^{K}\big(m_j(\theta) - \mu_{\mathrm{ANN}}(\theta)\big)^2}, \qquad \mathrm{WSPP}_{\mathrm{sur}}(\theta) = [\max(\mu(\theta),0)]^2 \qquad (23)$$

The use of an ensemble of networks to estimate the predictive uncertainty is inspired by *deep ensembles* [46]. Recent work shows, however, that the uncertainty thus obtained can be underestimated as the number of networks grows and that a posteriori calibration may be necessary [47]. We account for this by systematically verifying the solutions by simulation (Section 3.6). The performance of the surrogate is measured on the 180 test points by R^2 and by the mean absolute error, and compared with that of each component taken individually. The importance of each input variable is evaluated by permutation [48]:

$$\Delta R_j^2 = R^2_{\mathrm{base}} - \frac{1}{B}\sum_{b=1}^{B} R^2\big(\theta^{(j,b)}\big) \quad (24)$$

where $\theta^{(j,b)}$ denotes the test set whose j-th variable was randomly permuted at draw b and $B = 30$. A variable is all the more important as its permutation degrades the coefficient of determination.

3.6 Module 4: Genetic Algorithm with an Uncertainty Margin

The surrogate is used as the fitness function of a real-coded genetic algorithm. The variables are $(\ln V, s)$, with D and H fixed. To avoid retaining a solution whose reliability is satisfactory only because of a surrogate error, the constraint is applied to an upper bound of WSPP. Since $\mathrm{WSPP} = \mu^2$, first-order propagation of the standard deviation σ (delta method) gives $\sigma_{\mathrm{WSPP}} \approx 2\mu\sigma$, whence

$$\mathrm{WSPP}_{\mathrm{r}}(\theta) = \mathrm{WSPP}_{\mathrm{sur}}(\theta) + \kappa \cdot 2\sqrt{\mathrm{WSPP}_{\mathrm{sur}}(\theta)}\,\sigma(\theta) \quad (25)$$

$$J(\theta) = \mathrm{LCOW}(V, s) + \lambda \max(0,\ \mathrm{WSPP}_{\mathrm{r}}(\theta) - \varepsilon) \qquad (26)$$

where κ sets the level of prudence (here $\kappa = 1$), ε is the reliability threshold of Equation (14), and λ is a penalty coefficient large enough that any infeasible solution is dominated by a feasible one. The LCOW is computed by Equation (13) with $\mathrm{WSPP}_{\mathrm{sur}}$. The population evolves through tournament selection, elitism, and blend crossover: for two parents x_a and x_b, the offspring is

$$x_{\mathrm{child}} = \alpha\, x_a + (1-\alpha)\, x_b, \qquad \alpha \sim U(-0.25, 1.25) \quad (27)$$

followed by a Gaussian mutation applied with a fixed probability, and then by clipping to the bounds. The values of these parameters are given in Table 2. Since the algorithm is stochastic, it is run for 10 different seeds; the dispersion of the solutions obtained measures its robustness. Each solution is then re-evaluated by the simulator on the validation sequences, which are independent of those used for training and optimization. This step guarantees that the final results

rest on simulation and not on the surrogate. Finally, the optimum is compared with that of an exhaustive reference grid evaluated directly by simulation.

3.7 Pareto Front and Sensitivity Analysis

The trade-off between cost and reliability is characterized by a Pareto front. A configuration θ is dominated if there exists a configuration θ' such that

$$\mathrm{LCOW}(\theta') \leq \mathrm{LCOW}(\theta) \qquad \text{et} \qquad \mathrm{WSPP}(\theta') \leq \mathrm{WSPP}(\theta) \quad (28)$$

with at least one strict inequality. The front is the set of nondominated configurations. It is obtained by evaluating on the surrogate 200,000 configurations drawn at random in Θ, which would be infeasible with the simulator. Fourteen points of the front, evenly distributed, are then verified by simulation on the validation sequences, and the discrepancy between prediction and simulation is reported. This representation gives the decision maker the range of possible choices rather than a single solution.

Finally, the sensitivity analysis covers 17 scenarios around the reference case. The physical scenarios vary the demand, the head, the persistence ρ, the frequency of overcast days (no overcast day and 20 %), and the threshold ε (1 % and 10 %). The economic scenarios vary the tank cost (±50 %), the PV generator cost (±30 %), and the discount rate (5 % and 12 %). Each is re-optimized on a grid evaluated by simulation, without going through the surrogate, so that the sensitivity analysis does not depend on the surrogate's error.

Table 2: Parameters of the hybrid chain

Module	Parameters	Equations
Module 1	Gradient Boosting: $M = 300$ trees, depth 3, $\nu = 0.05$, subsample 0.8; 4 inputs	Eq. (15)-(17)
Module 2	$n_t = 4$ types; $\pi = (0.40, 0.25, 0.25, 0.10)$; $\rho = 0.5$; $N_R = 25$ sequences of 365 d	Eq. (18)-(20)
Data	$N = 900$ points (720 / 180); $V \in [20, 1500]$ m^3 (log); $s \in [0.5, 2.2]$; $D \in [200, 800]$ m^3/d; $H \in [20, 80]$ m	Eq. (21)
Module 3	$K = 5$ ANNs (2 × 48 neurons, tanh); Gradient Boosting: 400 trees, depth 3, $\nu = 0.05$; ½–½ weights; $B = 30$	Eq. (22)-(24)
Module 4	Population 60; 70 generations; elitism 2; tournament size 3; mutation: probability 0.3, standard deviation 4 % of the range; $\kappa = 1$; $\lambda = 10^3$; $\varepsilon = 5$ %; 10 seeds	Eq. (25)-(27)
Pareto	200,000 configurations; 14 points verified by simulation	Eq. (28)

4 CASE STUDY: THE SAKAL SITE

4.1 Site and Measured Data

The selected site is Sakal, a rural locality in northwestern Senegal, in the Louga region, close to the administrative boundary with the Saint-Louis region (Fig. 3). The climate is Sahelian; the approximate coordinates are 15.25° N and 16.47° W.

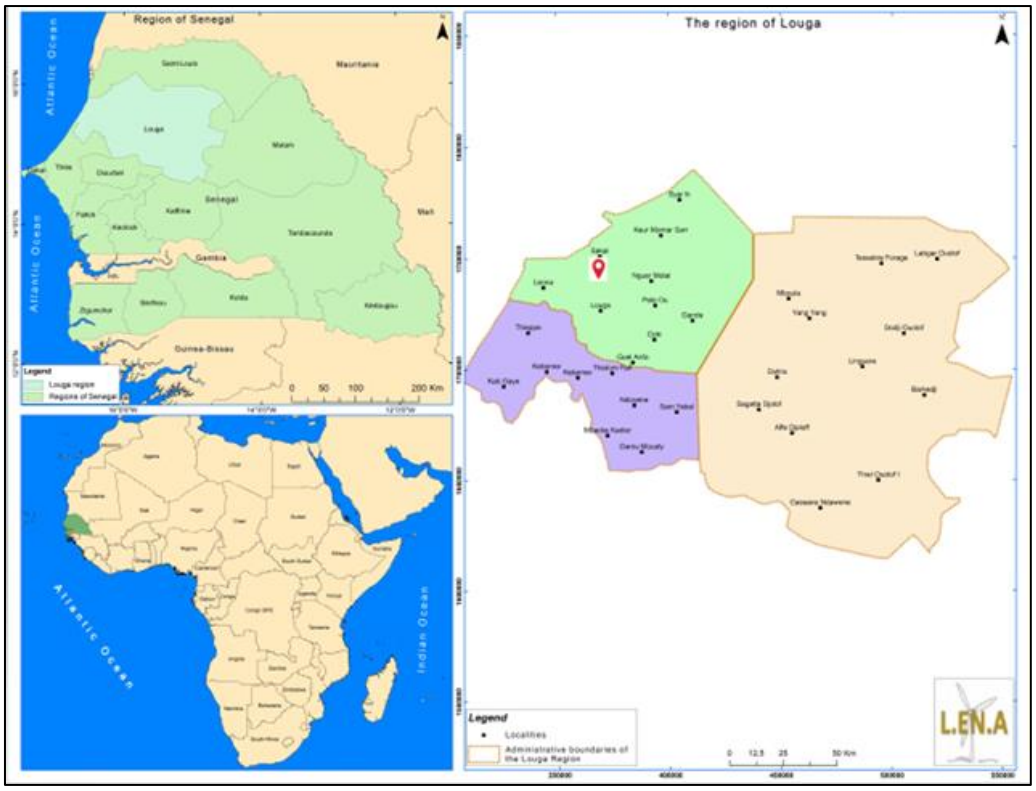


Figure 3: Location of the Sakal site in the Louga region (Senegal).

The data are the electrical power of a PV generator and the irradiance, measured at a 15-min time step (96 points per day) for three days in 2020: 15 January, 15 May, and 15 August. The maximum measured power is 16.3 kW; we therefore take as the reference generator a peak power $P_{c,ref} = 16$ kWp. These three days cover a clear day (May), a variable-sky day (August), and a cloudy day (January): Table 1 gives their characteristics.

Figure 4 presents the measurements. The clear day (May) gives a smooth power curve, with a plateau at about 15.5 kW from 10:00 to 16:00. The variable-sky day (August) alternates cloud passages and clear intervals, with abrupt drops in power. The cloudy day (January) falls to 5 kW at times and reaches 14.9 kW only during a clear spell. The power follows the irradiance very closely, which justifies the scaling (1).

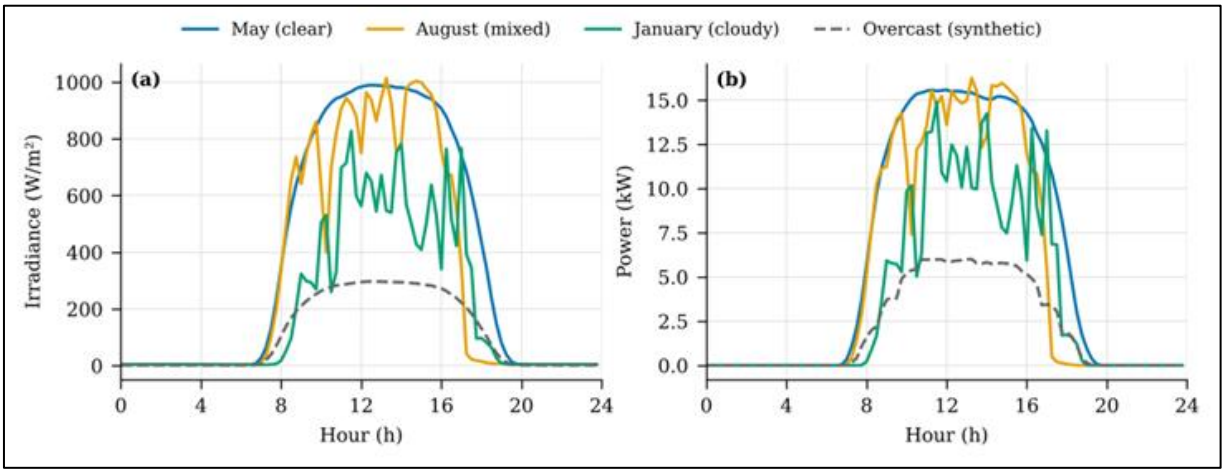


Figure 4: (a) Irradiance and (b) power of the reference generator for the three measured days of 2020 and for the synthetic overcast day (dashed).

Table 3: Measured days and synthetic typical day

Day type	Measurement	H (kWh/m²)	E (kWh)	P_{max} (kW)	Frequency π
Clear	15 May	9.03	143.2	15.6	0.40
Mixed	15 August	7.48	120.7	16.3	0.25
Cloudy	15 January	4.86	87.9	14.9	0.25
Overcast	synthetic	2.7	50.1	6.0	0.10

Note: H: daily irradiation; E: energy produced by the reference generator. The overcast day is constructed by reducing the irradiance of the clear day to 30 % of its value (and then applying Module 1); it is not measured.

4.2 Reference-Case Assumptions

The measured days do not include a heavily overcast day, yet it is these days that size the storage. A fourth "overcast" type is therefore added synthetically (Table 3). The type frequencies (π) and the persistence ($\rho = 0.5$) are working assumptions, tested in Section 5.8. The reference demand is $D = 400$ m^3/d (for indication, of the order of 13,000 to 20,000 users at 20 to 30 L/d per person), distributed according to the three-peak profile of Equation (6) (morning 7:00, midday 12:30, evening 18:30), and the total dynamic head is $H = 40$ m. The parameters of the conversion chain and the unit costs are those of Table 4. The tank cost follows a power law with exponent $b = 0.8 < 1$, consistent with the economies of scale observed for storage [36]. The costs are typical sizing values, not local quotations: the economic conclusions are therefore conditional on these values, which the sensitivity analysis quantifies.

Table 4: Parameters of the reference case and of the hybrid chain

Block	Values
System	$D = 400$ m^3/d; $H = 40$ m; $P_{c,ref} = 16$ kWp; $\eta_{nom} = 0.45$; $\kappa = 0.4$; $x_{min} = 0.08$; $v_0 = 0.5$; $\Delta t = 15$ min
Costs	$c_{pv} = 900$ USD/kWp; $c_p = 300$ USD/kW; $C_{fix} = 12{,}000$ USD; tank $a\,V^b$ with $a = 1200$ and $b = 0.8$ (V in m^3); operation: 1.5 % (PV), 3 % (pump), 1 % (tank) per year; pump replacement at 10 years (40 % of the cost); $r = 8$ %; $N = 20$ years; $\varepsilon = 5$ %
Module 1	Gradient Boosting: $M = 300$ trees, depth 3, $\nu = 0.05$, subsample 0.8
Module 2	$n_t = 4$; $\pi = (0.40, 0.25, 0.25, 0.10)$; $\rho = 0.5$; $N_R = 25$ sequences of 365 d
Module 3	$N = 900$ (720 / 180); $V \in [20{,}1500]$ m^3, $s \in [0.5{,}2.2]$, $D \in [200{,}800]$ m^3/d, $H \in [20{,}80]$ m; $K = 5$ networks (2 × 48, tanh); Gradient Boosting: 400 trees, depth 3, $\nu = 0.05$
Module 4	population 60; 70 generations; elitism 2; tournament size 3; mutation: probability 0.3, standard deviation 4 % of the range; $\kappa = 1$; $\lambda = 10^3$; 10 seeds
Pareto	200,000 configurations; 14 points verified by simulation

4.3 Implementation

The simulator, the learning chain, and the optimization are programmed in Python (*NumPy*, *SciPy*, *scikit-learn* [41]). A 25-year simulation at a 15-min time step, for one configuration (V, s), takes about 17 ms when 300 configurations are evaluated in batches. The complete computation of the study (reference grid, 900-simulation dataset, surrogate, 10 runs of the genetic algorithm, 200,000 front evaluations, 17 scenarios) takes about six minutes on a two-core computer. The results are presented in Section 5.

5 RESULTS AND DISCUSSION

5.1 Power Model

In leave-one-day-out cross-validation, the *Gradient Boosting* model reaches R^2 = 0.987 (January), 0.994 (May), and 0.994 (August), with an RMSE of 567, 532, and 489 W, i.e., 3 to 4 % of the reference power. Linear regression gives 753, 607, and 413 W. The gain is therefore 25 % in January and 12 % in May, but the model is worse than the linear baseline in August (+19 %). With only three days, this result does not allow a conclusion on the general superiority of the model;

it shows that it correctly reproduces the irradiance–power relationship, with a clear advantage on variable-sky or cloudy days, which are the ones that constrain the storage.

5.2 Without a Tank, Increasing PV Power Is Not Enough

With a 5 m^3 tank, the deficit remains at 56.2 % for 16 kWp (s = 1) and falls only to 52.5 % for four times this power (s = 4). The reason is structural: 25 % of the demand falls outside the pump operating hours on a clear day, and 40 % on a variable-sky day, and no additional power can serve this share. Multiplying the panels without a tank merely increases the curtailed energy. The tank is therefore not an accessory of the pump: it is what shifts the pumped water toward the consumption hours.

5.3 Techno-Economic Optimum

Figure 5a shows the levelized cost of water on the reference grid (34 volumes × 33 values of s); the gray zone corresponds to the sizings that do not satisfy WSPP ≤ 5 %. After local refinement, the optimum is a tank of $V^* = 225$ m^3, i.e., 0.56 day of autonomy, associated with a generator of $s^* = 1.45$ ($P_c = 23.2$ kWp), for an LCOW of 0.108 USD/m^3 and a deficit of 4.9 %. The optimum is a plateau rather than a point: all sizings with V between 205 and 240 m^3 and P_c between 20.8 and 29.6 kWp are within 2 % of the minimum cost. On independent climate sequences, the deficit is 4.4 % and the LCOW 0.107 USD/m^3, which confirms that the optimum does not depend on the sequences used to optimize it. The pump delivers 138,800 m^3/yr for a demand of 146,000 m^3/yr, and 42 % of the pumpable energy is curtailed: the excess power is the counterpart of a small tank.

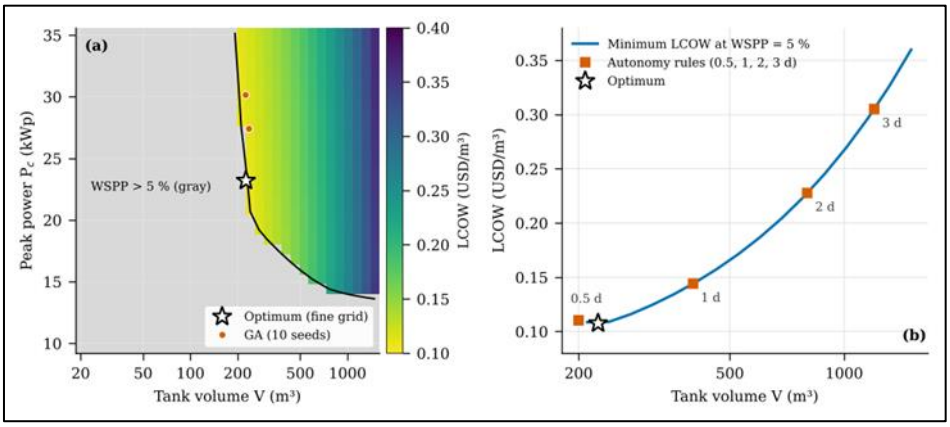


Figure 5: (a) LCOW as a function of tank volume and peak power; the black line is the limit WSPP = 5 %. (b) Minimum LCOW at WSPP = 5 % as a function of volume, with autonomy rules from 0.5 to 3 days.

5.4 Comparison with Autonomy Rules

Table 5 compares the optimum with the common rules expressed in days of autonomy [8]. For each rule, the volume is set to $n_j D$ and the power is adjusted to the minimum that satisfies WSPP ≤ 5 %. One day of autonomy raises the cost of water by 34 %, two days by 112 %, and three days by 183 %; an autonomy of a quarter of a day is infeasible for any power tested. The three-day rule [8] corresponds to 1200 m^3: the optimum reduces the storage by 81 % and the cost of water by 65 %, while increasing the PV power by 68 % (from 13.8 to 23.2 kWp). This conclusion is in line with, though stronger in magnitude than, that of an integrated sizing method that reduces the required storage by 50 % compared with the traditional method [9].

Table 5: Optimum and autonomy rules (WSPP ≤ 5 %, D = 400 m^3/d, H = 40 m)

Sizing	V (m^3)	P_c (kWp)	LCOW (USD/m^3)	Difference from minimum	Autonomy (d)
Optimum	225	23.2	0.108	—	0.56
0.5-day rule	200	31.6	0.110	+2 %	0.50
1-day rule	400	17.0	0.144	+34 %	1.0
2-day rule	800	14.3	0.228	+112 %	2.0
3-day rule	1200	13.8	0.305	+183 %	3.0

The cost breakdown explains this result (Fig. 6a). At the optimal point, the tank accounts for 62 % of the LCOW, the PV generator 14 %, operation 10 %, the fixed cost 8 %, and the pump 5 %. When the volume rises to 1500 m^3, the tank share reaches 85 %. Under the adopted costs, an additional kilowatt-peak costs far less than a cubic meter of storage removed, hence the substitution of storage by power: to meet the 5 % threshold, the required power goes from 27.5 kWp at 211 m^3 to 16.3 kWp at 462 m^3 and to 13.6 kWp at 1500 m^3. This trade-off, however, depends directly on the cost ratio (Section 5.8).

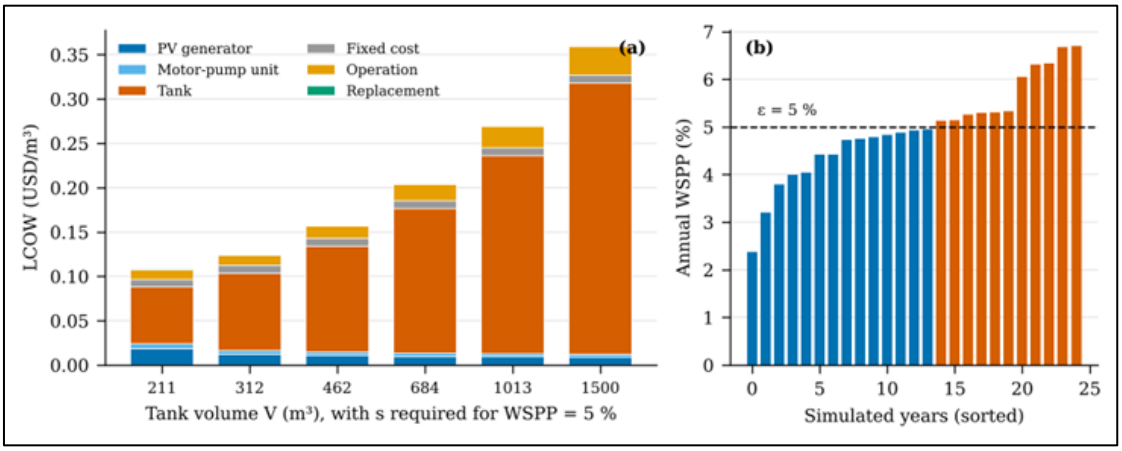


Figure 6: (a) LCOW breakdown as a function of volume, with the minimum power that satisfies WSPP = 5 %. (b) Annual deficit of the 25 simulated years at the optimum, sorted.

5.5 Interannual Variability

The constraint applies to the mean deficit of the 25 simulated years. At the optimum, the annual deficit has a mean of 4.9 % and a standard deviation of 1.0 point; its 95th percentile is 6.6 % and its maximum 6.7 % (Fig. 6b). Eleven years out of 25 (44 %) therefore exceed 5 %. Sizing on the mean is consistent with the chosen indicator, but a decision maker who requires the threshold to be met every year will have to accept a larger tank or higher power: this is one more reason to treat the constraint in probabilistic form, which the hybrid chain allows.

5.6 Performance of the Surrogate and of the Genetic Algorithm

On the 180 test points, the hybrid surrogate reaches $R^2 = 0.972$ and an MAE of 0.027 on WSPP, against 0.964 and 0.031 for the networks alone, and 0.960 and 0.030 for *Gradient Boosting* alone (Fig. 7a). Hybridization thus slightly improves each component. By permutation, the tank volume is the most influential parameter (drop in R^2 of 1.01), ahead of the demand (0.60), the power (0.43), and the head (0.33): this confirms, from the learning standpoint, that the volume is the primary lever of reliability.

The error is not uniform. On the 45 test points whose simulated deficit is below 10 %, the MAE is 0.018 and the bias +0.017: the surrogate overestimates the deficit near the threshold and thus errs on the side of caution. On the 14 front points verified by simulation (reference case), the mean absolute error is 0.018 overall, but 0.005 for the configurations whose simulated deficit is below 10 %. Conversely, in the transition zone (simulated deficit of 10 to 35 %), the surrogate underestimates the deficit by 0.03 to 0.05 (Fig. 7b). The error on the LCOW remains below 0.007 USD/m^3. The surrogate therefore does not remove the need for verification by simulation, which is maintained for all retained solutions.

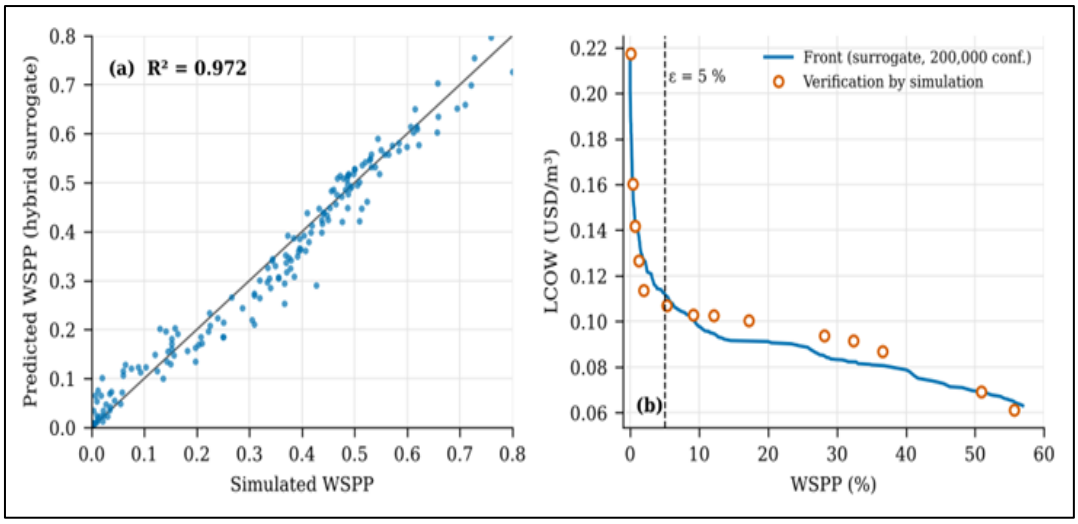


Figure 7: (a) WSPP predicted by the hybrid surrogate versus WSPP simulated on the 180 test points. (b) Cost–reliability Pareto front obtained with the surrogate and points verified by simulation.

Over the 10 seeds, the genetic algorithm converges 8 times to $V = 234.5$ m^3 and $P_c = 27.4$ kWp, and 2 times to $V \approx 223.6$ m^3 and $P_c \approx 30.2$ kWp; these two solutions belong to the plateau of Fig. 5a. Verified by simulation on independent sequences, they give a deficit of 2.4 to 2.7 % and an LCOW of 0.112 USD/m^3. The solutions are therefore cautious: the actual deficit is half the threshold, at the price of an LCOW about 3 % higher than that of the grid optimum (0.1086 USD/m^3 on the same sequences). This caution has two origins, the uncertainty margin $\kappa\sigma$ and the positive bias of the surrogate near the threshold; we did not separate them. It is the price of a guarantee against the surrogate's error.

Regarding computation time, the benefit of the surrogate must be qualified. For a single case (D and H fixed), a grid of 1122 simulations is sufficient and takes about 25 s: the surrogate brings no gain. Its value appears when the demand and the head vary (a dataset of 900 simulations covering the four parameters) and for exploring the front: a 25-year simulation takes 17.5 ms per configuration (in batches of 300), i.e., about one hour for 200,000 configurations, against less than one second for the surrogate (≈ 4000 times faster), the 900-simulation training set representing only about fifteen seconds.

5.7 Cost-Reliability Pareto Front

The front (Fig. 7b) gives the range of trade-offs. It is steep on the low-deficit side: from a deficit of 5 % (0.108 USD/m^3) to 1.2 %, the cost increases by 18 % (0.127 USD/m^3); at 0.7 %, by 32 %; at 0.4 %, by 49 %; at 0.1 %, by 102 % (0.217 USD/m^3). On the high-deficit side, the gain is small: accepting a 17 % deficit reduces the cost by only 7 % (0.100 USD/m^3). The 5 % threshold thus lies at the knee of the front, which justifies, a posteriori, the choice of ε for the reference case; a stricter threshold entails a rapidly increasing cost of water.

5.8 Sensitivity Analysis

Figure 8 summarizes the 17 scenarios, re-optimized on a finer grid than that of Section 5.3 (46 volumes from 80 to 900 m^3, 51 values of s) so that the optimal volume can vary; the volume step remains 5.6 %. The reference case gives $V = 235$ m^3 and $P_c = 21.1$ kWp for 0.108 USD/m^3, a point on the plateau of Section 5.3.

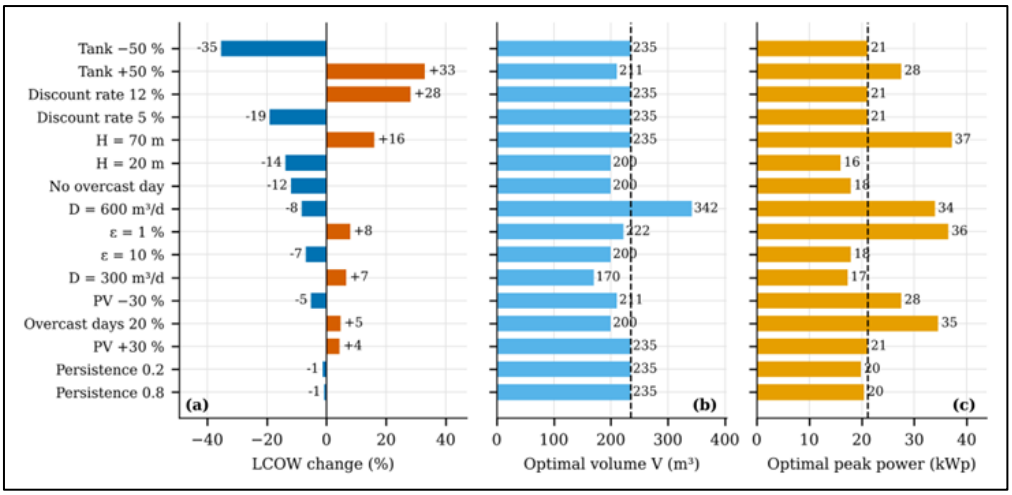


Figure 8: Sensitivity of the optimum to 17 scenarios: (a) variation of the LCOW relative to the reference, (b) optimal volume, (c) optimal peak power. Dashed lines indicate the reference.

The clearest result is the stability of the volume: across the 17 scenarios, the optimal autonomy remains between 0.50 and 0.59 day of demand. The volume is proportional to the demand ($V = 170$, 235, and 342 m^3 for $D = 300$, 400, and 600 m^3/d, i.e., 0.57 d in all three cases), and it varies little with the head, the tank cost (211 m^3 at +50 %, 235 m^3 at −50 %), the persistence, or the discount rate. This is explained by the role of the tank: it covers the share of the demand that falls outside the pumping hours (25 to 40 % of D), and this share is structural; multi-day autonomy, for its part, is bought with PV power, which is cheaper under the adopted assumptions. The parameters that do shift the optimum are, on the other hand, the power: it goes from 16 to 37 kWp when the head goes from 20 to 70 m, and from 18 to 35 kWp when overcast days go from 0 to 20 % of days. The frequency of overcast days, which is an assumption of the present work, is therefore one of the parameters to be established first from a local series.

On cost, the determinants are economic before being climatic: the tank cost (−35 % and +33 % for $\mp$50 %), the discount rate (−19 % at 5 %, +28 % at 12 %), and the head (−14 % at 20 m, +16 % at 70 m) weigh far more than the persistence (≈ −1 %) or the PV generator cost (−5 % and +4 % for $\mp$30 %). The reliability threshold changes the cost from +8 % (ε = 1 %) to −7 % (ε = 10 %). The qualitative conclusions (a volume close to half a day, storage dominating the cost, the advantage of power over autonomy) are therefore robust to the assumptions tested, whereas the absolute value of the LCOW depends strongly on the unit costs, which must be confirmed by local quotations.

5.9 Scope and Limitations of the Results

The results hold for the assumptions of Section 4 and should be read with four reservations. First, the power model is trained on three measured days, supplemented by a synthetic overcast day; the frequencies of the day types and the persistence are assumptions, among which the frequency of overcast days weighs on the optimal power (Section 5.8). Second, the unit costs and the demand profile are not local data: the ranking of the sizings is robust, but the absolute LCOW depends mostly on the tank cost. Third, the total head is assumed constant and the degradation of the generator and of the pump is neglected. Finally, the surrogate underestimates the deficit in the transition zone and therefore does not replace verification by simulation, which is maintained for all retained solutions.

6 CONCLUSION

This paper quantified the influence of the tank volume on the techno-economic sizing of a battery-less PV pumping system, using a hybrid chain that combines a *Gradient Boosting* power model, a climate-sequence generator based on typical day types and a Markov chain, a hybrid surrogate, and a genetic algorithm with an uncertainty margin, with systematic verification by simulation. Applied to Sakal (Louga, Senegal), for a demand of 400 m^3/d, a head of 40 m, and a tolerated deficit of 5 %, it leads to the following conclusions.

First, the tank is indispensable: without it, the deficit exceeds 52 % regardless of the installed power, because a significant share of the demand falls outside the pumping hours. Second, the optimum corresponds to a 225 m^3 tank (0.56 day of autonomy) and 23 kWp, i.e., 0.108 USD/m^3; it is a plateau, not an isolated point. Third, autonomy rules of one to three days raise the cost of water by 34 to 183 %: storage dominates the cost (62 % of the LCOW at the optimum), and under the adopted costs it is more economical to replace it with PV power. Fourth, the hybrid surrogate reproduces the deficit with $R^2 = 0.97$ and makes it possible to explore 200,000 configurations to trace the cost–reliability front, whose knee lies near 5 %.

These results have limitations that must be kept in mind. The power model relies on three measured days, to which a synthetic overcast day was added; the frequencies of the day types, the persistence, the demand profile, and the unit costs are assumptions, not local data; the total head is assumed constant; module degradation and pump aging are neglected; finally, a single site is studied. The surrogate underestimates the deficit in the intermediate zone of the front, and the 25 simulated years exceed the threshold in 44 % of cases.

Future work should replace the assumptions with local data (a multi-year irradiance series for Sakal, a record of the village's consumption, tank quotations), calibrate the surrogate's uncertainty (for example by conformal regression), and extend the method to other Sahelian sites and to a variable dynamic water level.

STATEMENTS ON COMPLIANCE WITH ETHICAL STANDARDS

Acknowledgments

The authors thank all those who supported this work, in particular the colleagues and institutions that facilitated access

Disclosure of conflict of interest

The authors declare no conflicts of interest regarding this manuscript.

REFERENCES

[1] WHO/UNICEF Joint Monitoring Programme (JMP), "Progress on household drinking water, sanitation and hygiene 2000–2024," UNICEF DATA, 2025. [Online]. Available: https://data.unicef.org/resources/jmp-report-2025

[2] N. Javanshir, S. Syri, A. Teräsvirta, and V. Olkkonen, "Abandoning peat in a city district heat system with wind power, heat pumps, and heat storage," Energy Reports, vol. 8, pp. 3051-3062, Nov. 2022, doi: 10.1016/j.egyr.2022.02.064.

[3] "Review of solar photovoltaic water pumping system technology for irrigation and community drinking water supplies," Renewable and Sustainable Energy Reviews, vol. 49, pp. 1084-1099, Sep. 2015, doi: 10.1016/j.rser.2015.04.083.

[4] A. Mazloumi, A. Poolad, M. S. Mokhtari, M. B. Altman, A. Y. Abdelaziz, and M. Elsisi, "Optimal Sizing of a Photovoltaic Pumping System Integrated with Water Storage Tank Considering Cost/Reliability Assessment Using Enhanced Artificial Rabbits Optimization: A Case Study," Mathematics, vol. 11, no. 2, p. 463, Jan. 2023, doi: 10.3390/math11020463.

[5] B. Abulkheir, E. Gouda, et al., "Optimal sizing of PV water pumping system for off-grid rural communities," Mansoura Engineering Journal, vol. 50, no 3, 2025. [Online]. Available: https://mej.researchcommons.org/home/vol50/iss3/17

[6] D. H. Muhsen, T. T. N. Khatib, and H. T. Haider, "A feasibility and load sensitivity analysis of photovoltaic water pumping system with battery and diesel generator," Energy Conversion and Management, vol. 148, pp. 287-304, Sep. 2017, doi: 10.1016/j.enconman.2017.06.007.

[7] A. H. Arab, F. Chenlo, and M. Benghanem, "Finding R&D Trends and Experts Made Simple | Scinapse," Accessed: Oct. 13, 2026. [Online]. Available: https://www.scinapse.io

[8] University of Arizona Cooperative Extension, "A mobile solar PV water pumping system," Publ. AZ1956, Tucson, AZ, USA, 2021. [Online]. Available: https://extension.arizona.edu/sites/default/files/2024-08/az1956-2021.pdf

[9] T. Hove and E. Mungofa, "Towards the Optimal Sizing of Solar-Powered Pump-Pipe-Storage Systems," in Proceedings of the ISES Solar World Congress 2019, Santiago, Chile: International Solar Energy Society, 2019, pp. 1-12. doi: 10.18086/swc.2019.34.02.

[10] Y. Bakelli, A. Hadj Arab, and B. Azoui, "Optimal sizing of photovoltaic pumping system with water tank storage using LPSP concept," Solar Energy, vol. 85, no. 2, pp. 288-294, Feb. 2011, doi: 10.1016/j.solener.2010.11.023.

[11] Y. Bakelli and A. Kaabeche, "Optimal size of photovoltaic pumping system using nature-inspired algorithms," International Transactions on Electrical Energy Systems, vol. 29, no. 9, p. e12045, Sep. 2019, doi: 10.1002/2050-7038.12045.

[12] I. M. Mahmoud, A. Ashraf, A. Hatem, S. O. Abdellatif, and H. A. Ghali, "Pre-sizing online tool for photovoltaic water pumping system: development, testing, validation, and evaluation," Discov Sustain, vol. 5, no. 1, p. 15, Feb. 2024, doi: 10.1007/s43621-024-00198-2.

[13] Z. Liang, A. L. Shorter, and A. G. Winter, "Quantifying the energetic cost tradeoffs of photovoltaic pumping systems for Sub-Saharan African smallholder farms," Irrig Sci, vol. 42, no. 1, pp. 149-161, Jan. 2024, doi: 10.1007/s00271-023-00872-9.

[14] R. J. Aguiar, M. Collares-Pereira, and J. P. Conde, "Simple procedure for generating sequences of daily radiation values using a library of Markov transition matrices," Solar Energy, vol. 40, no. 3, pp. 269-279, Jan. 1988, doi: 10.1016/0038-092X(88)90049-7.

[15] D. El-Shahat, A. Tolba, M. Abouhawwash, and M. Abdel-Basset, "Machine learning and deep learning models based grid search cross validation for short-term solar irradiance forecasting," Journal of Big Data, vol. 11, no. 1, Dec. 2024, doi: 10.1186/s40537-024-00991-w.

[16] M. Abdelsattar, M. A. Ismeil, M. M. A. A. Zayed, A. Abdelmoety, and A. Emad-Eldeen, "Assessing Machine Learning Approaches for Photovoltaic Energy Prediction in Sustainable Energy Systems," IEEE Access, vol. 12, pp. 107599-107615, 2024, doi: 10.1109/ACCESS.2024.3437191.

[17] J. H. Holland, Adaptation in Natural and Artificial Systems. Cambridge, MA, USA: MIT Press, 1992. Accessed: Oct. 13, 2026. [Online]. Available: https://mitpress.mit.edu/9780262082136/adaptation-in-natural-and-artificial-systems/

[18] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," IEEE Transactions on Evolutionary Computation, vol. 6, no. 2, pp. 182-197, Apr. 2002, doi: 10.1109/4235.996017.

[19] Y. Jin, "Surrogate-assisted evolutionary computation: Recent advances and future challenges," Swarm and Evolutionary Computation, vol. 1, no. 2, pp. 61-70, Jun. 2011, doi: 10.1016/j.swevo.2011.05.001.

[20] "Efficient Global Optimization of Expensive Black-Box Functions | Journal of Global Optimization | Springer Nature Link". Accessed: Oct. 13, 2026. [Online]. Available: https://link.springer.com/article/10.1023/A:1008306431147

[21] Y. Jin, H. Wang, T. Chugh, D. Guo, and K. Miettinen, "Data-Driven Evolutionary Optimization: An Overview and Case Studies," IEEE Transactions on Evolutionary Computation, vol. 23, no. 3, pp. 442-458, Jun. 2019, doi: 10.1109/TEVC.2018.2869001.

[22] "Machine Learning-Based Surrogate Modeling for Urban Water Networks: Review and Future Research Directions - Garzón - 2022 - Water Resources Research - Wiley Online Library". Accessed: Oct. 13, 2026. [Online]. Available: https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2021WR031808

[23] "Feasibility-guided evolutionary optimization of pump station design and operation in water networks | Scientific Reports". Accessed: Oct. 13, 2026. [Online]. Available: https://www.nature.com/articles/s41598-025-17630-w

[24] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," The Annals of Statistics, vol. 29, no. 5, pp. 1189-1232, 2001.

[25] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," Neural Networks, vol. 2, no. 5, pp. 359-366, Jan. 1989, doi: 10.1016/0893-6080(89)90020-8.

[26] "Photovoltaic Water Pumping Systems | Springer Nature Link". Accessed: Oct. 13, 2026. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-031-50791-5_2

[27] C. Soenen et al., "Comparison of Tank and Battery Storages for Photovoltaic Water Pumping," Energies, vol. 14, no. 9, p. 2483, Apr. 2021, doi: 10.3390/en14092483.

[28] T. Vezin et al., "Borehole water level model for photovoltaic water pumping systems," Applied energy, vol. 258, p. 114080, 2020.

[29] S. Meunier et al., "Influence of The Temporal Resolution of The Water Consumption Profile on Photovoltaic Water Pumping Systems Modelling and Sizing," in 2018 7th International Conference on Renewable Energy Research and Applications (ICRERA), Oct. 2018, pp. 494-499. doi: 10.1109/ICRERA.2018.8566828.

[30] "Effect of irradiance data on the optimal sizing of photovoltaic water pumping systems," Satie. Accessed: Oct. 13, 2026. [Online]. Available: https://satie.ens-paris-saclay.fr/en/node/940

[31] E. Skoplaki and J. A. Palyvos, "On the temperature dependence of photovoltaic module electrical performance: A review of efficiency/power correlations," 2009, doi: 10.1016/j.solener.2008.10.008.

[32] S. Meunier et al., "A validated model of a photovoltaic water pumping system for off-grid rural communities," Applied Energy, vol. 241, pp. 580-591, May 2019, doi: 10.1016/j.apenergy.2019.03.035.

[33] "Optimal Performance of Photovoltaic-Powered Water Pumping System". Accessed: Oct. 14, 2026. [Online]. Available: https://www.mdpi.com/2227-7390/11/3/731

[34] A. Młyńska and T. Bergel, “Research on the Hourly Water Consumption Structure in Rural and Suburban Household,” J. Ecol. Eng., vol. 23, no. 9, pp. 219-230, Sep. 2022, doi: 10.12911/22998993/151926.

[35] I. K. Okakwu, A. S. Alayande, T. O. Ajewole, and O. E. Olabode, “Techno-economic analysis of utilizing a directly-coupled photovoltaic water pumping system for domestic application in Nigeria,” Mehran Univ. res. j. eng. technol., vol. 42, no. 2, p. 197, Apr. 2023, doi: 10.22581/muet1982.2302.20.

[36] A. Parker, P. Cruddas, N. Rowe, R. Carter, and J. Webster, “Tank costs for domestic rainwater harvesting in East Africa,” Proceedings of the Institution of Civil Engineers - Water Management, vol. 166, no. 10, pp. 536-545, Nov. 2013, doi: 10.1680/wama.11.00113.

[37] N. Elmalı and C. Altın, “Machine Learning-Driven Rapid Optimization of Solar Power Plant Sizing Using HOMER-Generated Synthetic Scenarios,” Sustainability, vol. 18, no. 12, p. 6364, Jun. 2026, doi: 10.3390/su18126364.

[38] Y. Alakoç Bozkurt, C. Altın, and T. Çay, “Deep Learning-Supported Hybrid Renewable Energy System Optimization,” Solar, vol. 6, no. 4, p. 47, Aug. 2026, doi: 10.3390/solar6040047.

[39] X. Zhang et al., “Prediction of photovoltaic power generation based on a hybrid model,” Front. Energy Res., vol. 12, p. 1411461, Jun. 2024, doi: 10.3389/fenrg.2024.1411461.

[40] W. Xu, Z. Wang, W. Wang, J. Zhao, M. Wang, and Q. Wang, “Short-Term Photovoltaic Output Prediction Based on Decomposition and Reconstruction and XGBoost under Two Base Learners,” Energies, vol. 17, no. 4, p. 906, Feb. 2024, doi: 10.3390/en17040906.

[41] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” MACHINE LEARNING IN PYTHON.

[42] D. Liang, L. Ma, P. Wang, Y. Li, and Y. Luo, “Adaptive Design of Solar-Powered Energy Systems Based on Daily Clearness State Evolution,” Energies, vol. 17, no. 10, p. 2328, May 2024, doi: 10.3390/en17102328.

[43] S. Li, H. Ma, and W. Li, “Typical solar radiation year construction using k-means clustering and discrete-time Markov chain,” Applied Energy, vol. 205, pp. 720-731, Nov. 2017, doi: 10.1016/j.apenergy.2017.08.067.

[44] M. D. McKay, R. J. Beckman, and W. J. Conover, “Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code,” Technometrics, vol. 21, no. 2, pp. 239-245, May 1979, doi: 10.1080/00401706.1979.10489755.

[45] P. Borisut and A. Nuchitprasittichai, “Adaptive Latin Hypercube Sampling for a Surrogate-Based Optimization with Artificial Neural Network,” Processes, vol. 11, no. 11, p. 3232, Nov. 2023, doi: 10.3390/pr11113232.

[46] B. Lakshminarayanan, A. Pritzel, and C. Blundell, “Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles,” in Advances in Neural Information Processing Systems, Curran Associates, Inc., 2017. Accessed: Oct. 17, 2026. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/9ef2ed4b7fd2c810847ffa5fa85bce38-Abstract.html

[47] S. Yang and K. Yee, “Towards Reliable Uncertainty Quantification via Deep Ensembles in Multi-output Regression Task,” Nov. 24, 2023, arXiv: arXiv:2303.16210. doi: 10.48550/arXiv.2303.16210.

[48] A. Fisher, C. Rudin, and F. Dominici, “All Models are Wrong, but Many are Useful: Learning a Variable’s Importance by Studying an Entire Class of Prediction Models Simultaneously,” Journal of Machine Learning Research, vol. 20, no. 177, pp. 1-81, 2019.