

Optimizing Photovoltaic Panel Quantity for Water Distribution Networks

Ürkmez, Mirhan; Kallesøe, Carsten Skovmose; Bendtsen, Jan Dimon; Leth, John-Josef

Published in:
IEEE Transactions on Control Systems Technology

DOI (link to publication from Publisher):
[10.1109/TCST.2025.3560173](https://doi.org/10.1109/TCST.2025.3560173)

Publication date:
2025

Document Version
Accepted author manuscript, peer reviewed version

[Link to publication from Aalborg University](#)

Citation for published version (APA):
Ürkmez, M., Kallesøe, C. S., Bendtsen, J. D., & Leth, J.-J. (2025). Optimizing Photovoltaic Panel Quantity for Water Distribution Networks. *IEEE Transactions on Control Systems Technology*, 33(5), 1727-1742.
<https://doi.org/10.1109/TCST.2025.3560173>

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal -

Take down policy

If you believe that this document breaches copyright please contact us at vbn@aub.aau.dk providing details, and we will remove access to the work immediately and investigate your claim.

Optimizing Photovoltaic Panel Quantity for Water Distribution Networks

Mirhan Ürkmez¹, Carsten Kallesøe², Jan Dimon Bendtsen³, *Member, IEEE*, and John Leth⁴

Abstract—The article introduces a procedure for determining an approximation of the optimal amount of photovoltaics (PVs) for powering water distribution networks (WDNs) through grid-connected PVs. The procedure aims to find the PV amount, minimizing the total expected cost of the WDN over the lifespan of the PVs. The approach follows an iterative process, starting with an initial estimate of the PV quantity and then calculating the total cost of WDN operation. To calculate the total cost of the WDN, we sample PV power profiles that represent the future production based on a probabilistic PV production model. Simulations are conducted assuming these sampled PV profiles power the WDN, and pump flow rates are determined using a control method designed for PV-powered WDNs. Following the simulations, the overall WDN cost is calculated. Since we lack access to derivative information, we employ the derivative-free Nelder-Mead method for iteratively adjusting the PV quantity to find an approximation of the optimal value. The procedure is applied for the WDN of Randers, a Danish town. By determining an approximation of the optimal quantity of PVs, we observe a 14.5% decrease in WDN costs compared to the scenario without PV installations, assuming a 25-year lifespan for the PV panels.

Index Terms—Photovoltaic (PV) panels, predictive control, simulation-based optimization, water distribution networks (WDNs).

I. INTRODUCTION

WATER distribution networks (WDNs) are complex infrastructures that provide drinkable water to consumers in a specific area. Since electricity consumption in WDNs constitutes around 7%–8% of the world's total energy consumption [1], numerous researchers have explored the problem of scheduling water pumps to minimize the operational costs of WDNs [2], [3]. One potential solution involves powering WDNs using renewable energy sources like photovoltaic (PV) panels, which are easy to install and maintain. Specifically, grid-connected PVs can address the intermittent energy production issue of PV panels by allowing the grid to power WDN when there is insufficient PV power. To implement this solution, one must determine the appropriate number of PV panels to install for a given WDN. Existing literature on this subject typically relies on simulating WDNs with pump scheduling methods to find an approximation of

the optimal PV quantity, aiming to minimize the total cost of WDNs over the lifespan of PVs [4], [5]. However, these pump scheduling methods are often non-adaptive, meaning they cannot adjust pump schedules based on available PV forecasts and electricity prices to optimize WDN operating costs. Therefore, if a facility plans to use adaptive methods for operating its PV-powered WDN, depending on PV panel quantity recommendations from methods that simulate the network with non-adaptive pump scheduling may yield unreliable suggestions. This article presents a procedure for the determination of an approximate value of the optimal quantity of PV panels for a specific WDN by optimizing the total cost of the network throughout the lifespan of PV panels through the use of an adaptive pump scheduling simulation.

To calculate the total cost of the WDN, it needs to be simulated using a pump scheduling method. The design of a pump scheduling method for WDNs is primarily defined by the considerable sizes of the networks and the presence of nonlinear network elements. Most of the available works in the literature can be classified based on their strategies for addressing the challenges posed by these nonlinearities and network sizes. Heuristic methods are used in some works to avoid running gradient-based optimization algorithms with complex models of the networks and nonlinear constraints. In [6], the hydraulic simulation software EPANET is employed to generate and validate feasible solutions within a heuristic approach. Both Ormsbee and Lingireddy [7] and Castro-Gama et al. [8] use a genetic algorithm (GA), with Ormsbee and Lingireddy [7] using pump operation times as decision variables and Castro-Gama et al. [8] use binary variables for pump opening and closing.

Some works have tried simplifying the network model, typically by relaxing the nonlinear pipe equations. In [9], for example, the approach involves approximating the pipe equations as linear equations and applying model predictive control (MPC). The same linearization strategy is used in [10] for a PV-powered WDN. Similarly, an economic MPC (EMPC) method is developed with a set of linear inequalities derived from the relaxation of the nonlinear pipe equations [11]. Another study, [12], adopts a different methodology by simplifying the network structure through node clustering. Each cluster is then represented by a single node, and a system model is developed using a deep neural network (DNN) structure based on this simplified configuration. Goryashko and Nemirovski [13] tackle the pump scheduling problem under uncertain water demand by introducing a linear programming

Received 10 April 2024; revised 19 August 2024 and 22 January 2025; accepted 17 March 2025. This work was supported by the Independent Research Fund Denmark (DFF). Recommended by Associate Editor V. Adetola. (Corresponding author: Mirhan Ürkmez.)

The authors are with the Department of Electronic Systems, Aalborg University, 9220 Aalborg, Denmark (e-mail: mu@es.aau.dk; csk@es.aau.dk; dimon@es.aau.dk; jjl@es.aau.dk).

Digital Object Identifier 10.1109/TCST.2025.3560173

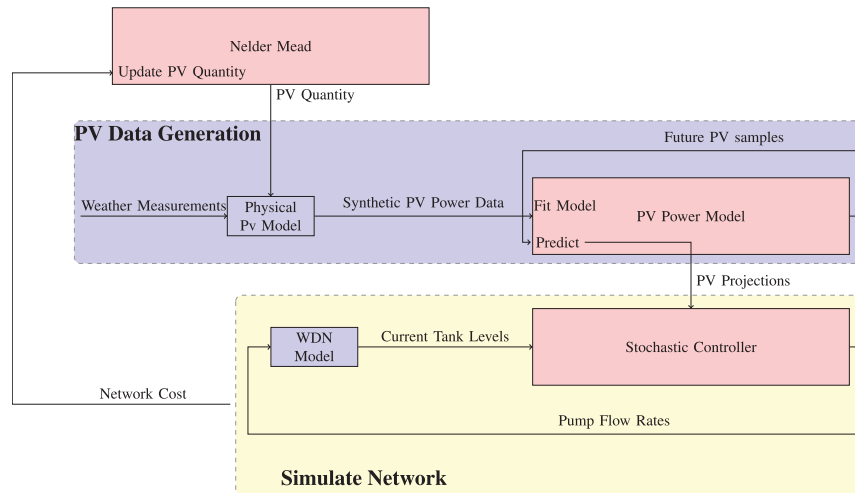


Fig. 1. Overview of PV quantity optimization method. The Nelder-Mead method iteratively updates the PV quantity. For each PV quantity, synthetic historical PV power data is generated using a physical PV model and weather data. The PV power model is fit to this data, and PV samples are generated. An approximate WDN model [see (21)] is simulated using these samples as actual PV power. A stochastic controller schedules pumps based on projected PV power from the PV model. The total cost, including grid power, PV installation, and maintenance, is calculated, and the PV quantity is updated via Nelder-Mead method.

framework and employing a robust methodology to handle the associated uncertainties. In [3], an EMPC is applied to a reduced model of the network. The EMPC formulation does not consider node pressure constraints. This omission is based on the assumption that the network has an elevated tank, and as long as the tank water levels are maintained above a certain threshold, the resulting pressure levels will remain within acceptable bounds. In this work, the same assumption is adopted. In recent works [14], [15], EMPC-based pump scheduling methods are proposed for WDNs powered by grid-connected PV panels. The main difference between the method given in [14] and the one in [15], which is also used in this article, is the cost functions used in the controller calculation. The method in [14] uses a deterministic cost function, while the one in [15] uses a stochastic cost function to account for potential future scenarios in PV production.

The use of PV panels in any system requires the ability to predict their power production due to the volatility in power production. The PV power prediction problem is studied extensively in the literature, with various prediction horizons and types of data such as weather measurements, historical power production data, and numerical weather predictions (NWP). NWP is employed alongside physical models of PVs to estimate power production using a closed-form equation, incorporating weather variables like radiance, temperature, and wind speed [16], [17]. Some studies categorize days as sunny, cloudy, or rainy using NWP and adjust the predictions accordingly [18], [19]. Some works solely rely on historical power data for their predictions. In [20], the use of NWP is avoided due to concerns about its availability in all locations. For short-term predictions, historical power data takes precedence as it is considered a more significant source of information, particularly for predictions within a few hours [21], [22]. In [23], a comparative analysis was conducted between an NWP-based physical model and several NWP-independent methods for predicting the day-ahead power values. The findings revealed that the non-NWP-based methods performed

better in predicting the power values around noon compared to the NWP-based physical model.

Researchers have been looking into how to figure out the right number of PV panels to keep costs down in various systems [24], [25], [26], [27], [28]. In the context of WDNs powered by PVs, studies fall into two main types: those using grid-connected PV panels and those using off-grid solar systems. For off-grid systems, the main challenge is finding the amount of solar power needed to meet the network's energy demands [29]. In [30], a system with an energy storage system (ESS) is considered, and the average yearly energy production of PVs is used to calculate the PV amount for WDNs. In [31], an approximation of the optimal PV size is determined using the firefly algorithm, with the constraint that daily PV energy output exceeds the daily energy required by water pumps to meet water demand, as the system did not have an ESS. In [32], the PV amount is determined to minimize the operational costs while maintaining a reliability threshold, measured as total delivered water divided by total water demand. For grid-connected systems, the challenge is a bit different. Now, the total energy from solar panels does not have to match the energy needed by WDNs because grid energy can help out. The goal is to find the right amount of PV panels to minimize costs over the life of the panels. In [5], an approximation of the optimal PV amount is found for a grid-connected PV-powered irrigation system by simulating the system with an irrigation scheduling method. Another study [4] looked at designing solar-powered WDNs and optimized the number of panels to minimize total operating costs using a predetermined pump scheduling strategy. It is worth noting that in both [4] and [5], the pump schedules were not updated during simulations to minimize operating costs of the WDN, leading to a significant reduction in simulation time. However, if a facility aims to minimize costs by installing a minimal number of PVs and implementing adaptive pump scheduling methods, the recommendations from these approaches may be unreliable, as they do not incorporate adaptive pump scheduling methods.

This article presents two primary contributions. The first contribution is offering a systematic approach to determine an approximation of the optimal amount of PV panels to be installed for the desired pumping stations in a given WDN. The objective is to find the PV panel quantity that optimizes the network's overall cost throughout the lifespan of the PV panels, encompassing installation, maintenance, and water pump usage costs. This objective is chosen to enhance the appeal of installing PV panels in WDNs. The calculation of the total cost for a given amount of PV panels is done by generating samples of PV power production with a new probabilistic PV power production model and simulating the network using these samples in conjunction with a stochastic predictive controller-based pump scheduling method introduced in [15] because of its ability to consider the uncertainties associated with PV power production when establishing pump schedules. Iterative updates to the PV panel quantities are performed using a derivative-free optimization algorithm to find an approximation of the optimal number of PV panels, minimizing the total cost. The choice of a derivative-free method stems from the unavailability of gradient information for the total WDN costs. This is due to the fact that the total cost depends on calculating pump schedules through an optimization problem. Unlike prior research focused on optimizing the quantity of PV panels, the approach proposed in this article involves using an adaptive pump scheduling method that tailors pump schedules based on the PV forecasts available and the electricity prices to minimize economic costs. Therefore, this method can provide recommendations on the quantity of PV panels for facilities that want to use adaptive pump scheduling methods in the future. The proposed method is implemented on a linear model representing a medium-sized Danish town's network, specifically the Randers network. The flowchart of the optimization procedure is given in Fig. 1.

The second contribution is introducing a new probabilistic PV power production model. Building upon the method introduced in [33] and refined in [15], the improvements include using a multiplicative correction term [see (4)] instead of an additive one given in [15] to better model the intraday fluctuations in PV power production due to weather conditions. Additionally, the process of sampling future PV power has been improved compared to the one in [15], where the probability densities of the additive correction terms have not been updated with the arrival of the new PV values. In this article, incoming PV values are used to update the densities of multiplicative correction terms. Finally, different stochastic models than those in [33] are used in this article to predict specific variables in the framework, which helps in the accurate modeling of PV data from locations where there are significant seasonal changes in solar power production.

The outline of the rest of the article is as follows. PV power production model is given in Section II. The model of the network is derived in Section III. The control method is explained in Section IV. The procedure for optimizing the PV amount is explained in Section V. The experimental results are presented in Section VI. The article is concluded in Section VII.

A. Notation

The time derivative of a function $f(t)$ is shown with $\dot{f}(t)$. The sign \odot is used to denote the elementwise multiplication of two matrices. The probability density function of a random variable X is shown with $f_X(\cdot)$. The conditional density of X given the value y of the random variable Y is written as $f_{X|Y}(\cdot|y)$. The Pontryagin difference is denoted with \ominus , where $A \ominus B = \{a|a + B \subseteq A\}$.

II. PV POWER PRODUCTION MODEL

In this section, the PV power production model is introduced, which is based on a method outlined in [33]. The model serves two main purposes. First, it is used to generate future PV profile samples for a given number of PVs, enabling the simulation of the WDN with these samples as its power source. Second, the model is used to create realizations of future PV power production with a maximum one-day horizon. This helps in approximating the expected value in the cost function [see (22a)] for the pump scheduling method. To fit the model parameters for the location where PVs are intended to be installed, the historical PV power production data is needed. However, assuming the availability of such historical data is not feasible, as the goal is to determine an approximation of the optimal PV installation without any PVs currently in place. To address this challenge, a synthetic historical PV power dataset is generated by utilizing historical weather measurements, such as irradiance, temperature, and wind speed. These measurements are inputted into physical models that approximate PV power output as a function of weather variables. With the synthetic historical PV power data in hand, one can construct the PV power production model and fit its parameters using this synthetic historical dataset. One might question the necessity of having a distinct model when physical models are already available. The reason is that those physical models will not work for the problem considered in this article. A model capable of creating PV samples for a long time into the future, one year for the simulations in this article, is required. However, the existing physical models require weather predictions for that duration, which is not available. Also, the model needs to be able to represent PV power probabilistically. This is necessary to generate realizations of future PV power with a one-day horizon, which are needed by the pump scheduling method.

A. Synthetic Power Data

To generate synthetic PV power data, the model presented in [16] is used, which is given by

$$P(G', T') = G' (P_{\text{STC}} + k_1 \ln(G') + k_2 \ln(G')^2 + k_3 T' + k_4 T' \ln(G') + k_5 T' \ln(G')^2 + k_6 T'^2) \quad (1)$$

where $k_i, i = 1, \dots, 6$, are constants to be estimated for each PV panel type, and G' and T' are the normalized irradiance and module temperature values, given by

$$G' = G/G_{\text{STC}} \quad (2a)$$

$$T' = T_{\text{mod}} - T_{\text{STC}} \quad (2b)$$

with $G_{\text{STC}} = 1000 \text{ W/m}^2$, and $T_{\text{STC}} = 25 \text{ }^\circ\text{C}$ being the irradiance and temperature values under standard conditions. P_{STC} is the power production under normal conditions, i.e., $G = G_{\text{STC}}, T_{\text{mod}} = T_{\text{STC}}$. The temperature of the PV module T_{mod} can be approximated using the Faïman model presented in [34] as

$$T_{\text{mod}} = T_{\text{amb}} + \frac{G}{\mu_0 + \mu_1 V_{\text{wind}}} \quad (3)$$

where T_{amb} is the ambient temperature, V_{wind} is the wind speed, and μ_0 and μ_1 are constant parameters. To estimate the parameters $k_1 - k_6, \mu_0, \mu_1$, a small number of PVs should be installed, and the irradiance, temperature, and power measurements must be taken for a while. Then, synthetic historical PV power data can be generated using historical weather measurements.

B. Pv Power Model

The challenge in modeling PV power generation is that it has both seasonal and daily patterns, along with intraday fluctuations. The sun's daily trajectory relative to PV panels changes throughout the year. Consequently, sunrise and sunset times, as well as the PV production at each hour of the day compared to other hours, undergo seasonal changes. The angle of sunlight also changes throughout the year, impacting the average daily PV production. Weather conditions contribute to the complexity by causing intraday fluctuations in PV production. In this section, a PV model capturing these characteristics is presented. The PV power production is modeled as

$$X_\eta = p_\eta Y_\eta \odot \delta_\eta \quad (4)$$

where $X_\eta \in \mathbb{R}^{N_{\text{pv}}}$ is the power output of the PVs at day η , $N_{\text{pv}} \in \mathbb{R}_{>0}$ is the number of data points in a day, $p_\eta \in \mathbb{R}$ is the optimal multiplier [see (6)], $Y_\eta \in \mathbb{R}^{N_{\text{pv}}}$ is the daily power profile, and $\delta_\eta \in \mathbb{R}^{N_{\text{pv}}}$ is the multiplicative correction term [see (9)]. Note that multiplicative correction terms are used in this article instead of the additive ones given in [15]. The purpose of the model is to enable day-ahead power forecasting and generate samples of PV profiles.

The daily power profile Y_η represents the change of the sun's position in a day and provides information about the relative amount of irradiation received at each time during the day. Since the sun's intraday position changes are very similar between consecutive days, Y_η is calculated using the exponentially weighted moving average (EWMA) method as

$$X'_{\eta-1} = \frac{1}{g(\eta-1)} X_{\eta-1}, \quad Y_\eta = \alpha X'_{\eta-1} + (1-\alpha) Y_{\eta-1} \quad (5)$$

where $g(\eta-1)$ is a smoothed form of the daily maximum power production $\max(X_{\eta-1})$. This smoothing is achieved by fitting a one-year periodic function to the daily maximum production data $\max(X_\tau)$, $\tau = 1, \dots, N_s$ of synthetic PV data using least squares, where N_s is the number of days in the synthetic dataset. An illustration of the function $g(\tau)$ along with the corresponding daily maximum power production is given in Fig. 2. The power profile Y_η is then obtained by using EWMA (5) with the normalized daily data $X'_{\eta-1} \in \mathbb{R}^{N_{\text{pv}}}$. The EWMA parameter $\alpha \in [0, 1]$ governs the influence of

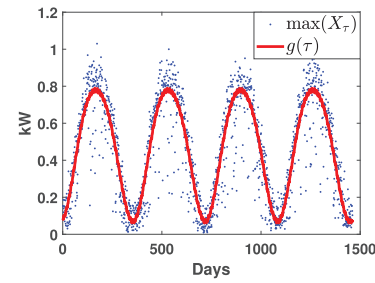


Fig. 2. Illustration of the $g(\tau)$ fit to the time series of the daily maximum power production $\max(X_\tau)$ of the synthetic PV data generated with $P_{\text{STC}} = 1 \text{ kW}$.

recent days relative to earlier ones in shaping Y_η . The rationale behind using normalized daily data X' instead of X in the computation of the next day's daily profile Y_η is to balance the influence of days characterized by high PV production and those with low PV production on the values of Y_η .

The optimal multiplier p_η is defined as

$$p_\eta = \arg \min_p \sum_{i=1}^{N_{\text{pv}}} (p Y_\eta^i - X_\eta^i)^2 = \frac{\sum_{i=1}^{N_{\text{pv}}} Y_\eta^i X_\eta^i}{\sum_{i=1}^{N_{\text{pv}}} Y_\eta^{i2}} \quad (6)$$

where Y_η^i (resp. X_η^i) is the i th coordinate of Y_η (resp. X_η). As it is the multiplier of Y_η such that $p_\eta Y_\eta$ is closest to the real production data X_η , it can be thought as representing the average daily weather condition which gets bigger values on clear days, and smaller values on cloudy days. Calculating the optimal multiplier p_η relies on having all the power values for the current day X_η accessible through (6). Consequently, to estimate the power values X_η before the day η , a probabilistic model for the optimal multiplier p_η is needed. Optimal multipliers are modeled as

$$p_\eta = (\gamma(\eta) + \epsilon_{p_\eta})^2 \quad (7)$$

where $\gamma(\eta)$ is the term representing the seasonal changes of optimal multipliers, and ϵ_{p_η} is the error term representing the daily deviations from the seasonal changes. Please note that the original model introduced in [33] did not include a seasonal term for the optimal multipliers, so it could not account for seasonal changes in PV production. The seasonal part $\gamma(\eta)$ is obtained by fitting a periodic function with a period of one year to the time series of the square root of optimal multipliers $\sqrt{p_\tau}$, $\tau = 1, \dots, N_s$ of the synthetic dataset using least squares. The error terms ϵ_{p_τ} , $\tau = 1, \dots, N_s$ are then obtained as $\epsilon_{p_\tau} = \sqrt{p_\tau} - \gamma(\tau)$. Since the optimal multiplier is a nonnegative variable, it is modeled as the square of a random variable so that when ϵ_{p_η} is sampled from its respective distribution, the resulting optimal variable p_η is always nonnegative. The error values ϵ_{p_η} are modeled using an autoregressive moving average model, ARMA(1, 1), as

$$\epsilon_{p_\eta} = \mu_p + \phi_p \epsilon_{p_{\eta-1}} + \theta_p \zeta_{p_{\eta-1}} + \zeta_{p_\eta} \quad (8)$$

where ζ_{p_η} and $\zeta_{p_{\eta-1}}$ are the forecast errors at step $\eta, \eta-1$, respectively and μ_p, ϕ_p, θ_p are the constants whose values are found by fitting the time series ϵ_{p_τ} , $\tau = 1, \dots, N_s$ to the

model. The error terms ζ_{p_η} are assumed to be coming from independent zero-mean normal distributions.

Multiplicative correction terms δ_η are defined as

$$\delta_\eta^i = \frac{X_\eta^i}{p_\eta Y_\eta^i}. \quad (9)$$

They capture the intraday changes in weather conditions. If the weather remains stable throughout a day, either clear or cloudy, correction terms δ_η stay around the value 1. On the other hand, if there is an inconsistency in the weather conditions, for example, a clear sky at time instance i on a cloudy day η , the correction term $\delta_\eta^i > 1$ attains a bigger value. The choice of multiplicative correction terms, as opposed to the additive ones used in [15], aims to create a correction term time series whose statistical properties, such as mean and variance, do not vary a lot in time so that it can be better modeled with a stationary process. When there is a change in weather conditions during high PV production, the impact on PV power production is more significant compared to periods of low production. For instance, a sudden cloud cover would cause a greater reduction in PV power production at noon than near sunset. In an additive correction term model, this would lead to larger correction term values at noon and smaller ones away from noon. However, in the multiplicative correction term model, the values of correction terms at noon and during other times are closer to each other. Similar to optimal multipliers, a probabilistic model for correction terms δ_η^i is necessary to estimate PV power X_η , since these terms are determined by the power values of the current day according to (9). The correction terms at night are insignificant since both the power production X_η and the daily power profile Y_η values are zero. Therefore, only the correction terms during the day are modeled as

$$\ln \delta_\eta^i = \mu_\delta + \phi_\delta \ln \delta_\eta^{i-1} + \zeta_{\delta_\eta}^i \quad (10)$$

where μ_δ, ϕ_δ are constant parameters and $\zeta_{\delta_\eta}^i$ are assumed to be coming from independent zero-mean normal distributions. Given that the correction terms δ_η^i , as defined in (9), are inherently positive, using the logarithmic function \ln to represent these correction terms in the model (10) ensures that the predictions for the correction terms will also remain positive. This model is developed solely based on the historical time series of the correction terms. However, when the previous day $\eta - 1$ ends and the daily power profile Y_η is obtained using (5), it tells more about δ_η which could be used to update the distributions of the $\zeta_{\delta_\eta}^i$ terms. Specifically, multiplicative correction terms should satisfy a certain equation which could be derived as

$$p_\eta = \frac{\sum_{i=1}^{N_{pv}} Y_\eta^i X_\eta^i}{\sum_{i=1}^{N_{pv}} Y_\eta^{i2}} = \frac{\sum_{i=1}^{N_{pv}} Y_\eta^i p_\eta Y_\eta^i \delta_\eta^i}{\sum_{i=1}^{N_{pv}} Y_\eta^{i2}} \quad (11)$$

$$\sum_{i=1}^{N_{pv}} Y_\eta^{i2} \delta_\eta^i = \sum_{i=1}^{N_{pv}} Y_\eta^{i2}.$$

It can be seen that the true values of the correction terms satisfy this equation by substituting δ_η^i with its definition (9) into (11). This substitution yields the definition of the optimal

multiplier (6), which means that the variable definitions in the PV model are consistent with (11). After the preceding day $\eta - 1$ and the acquisition of the daily power profile Y_η , the correction terms δ_η^i remain unknown. In the absence of (11), the correction term model (10) would be used for estimating these terms. However, incorporating the information from (11) with the correction term model (10) following the arrival of Y_η allows for a more accurate estimation of the correction terms. With this information, variables $\zeta_{\delta_\eta}^i$ are no longer independent of each other. Therefore, when the value of the daily power profile Y_η becomes available, and (11) can be used, the initial distributions of $\zeta_{\delta_\eta}^i$, which are independent zero-mean normal distribution, cannot be used to sample or estimate them. Instead, the probability densities of the $\zeta_{\delta_\eta}^i$ terms are updated at the beginning of the day η using (11) as

$$f_{\zeta_{\delta_\eta}^{sr}, \zeta_{\delta_\eta}^{sr+1}, \dots, \zeta_{\delta_\eta}^{sd} | \sum_{i=1}^{N_{pv}} Y_\eta^{i2} \delta_\eta^i} \left(Z_{\delta_\eta}^{sr}, Z_{\delta_\eta}^{sr+1}, \dots, Z_{\delta_\eta}^{sd} \middle| \sum_{i=1}^{N_{pv}} Y_\eta^{i2} \right) \propto \begin{cases} f_{\zeta_{\delta_\eta}^{sr}}(Z_{\delta_\eta}^{sr}), \dots, f_{\zeta_{\delta_\eta}^{sd}}(Z_{\delta_\eta}^{sd}), & \sum_{i=1}^{N_{pv}} Y_\eta^{i2} \delta_\eta^i(Z) = \sum_{i=1}^{N_{pv}} Y_\eta^{i2} \\ 0, & \text{otherwise} \end{cases} \quad (12)$$

where sr, sd are the time indices for sunrise and sundown, respectively, and $\delta_\eta^i(Z)$, $i = sr, \dots, sd$, are the correction terms obtained by substituting the sampled values $Z_{\delta_\eta}^i$, $i = sr, \dots, sd$, in place of the corresponding $\zeta_{\delta_\eta}^i$ in the correction error term model (10). This probability update essentially results from the combination of the general information concerning the correction terms (10) and the new information (11) obtained at the beginning of the day η . Before obtaining the daily power profile Y_η , the occurrence of any sequence of $Z_{\delta_\eta}^i$, $i = sr, \dots, sd$ had a positive probability, since $\zeta_{\delta_\eta}^i$ terms are assumed to be coming from independent zero-mean normal distributions. Once Y_η values become available, only those specific sequences of $Z_{\delta_\eta}^i$ that yield $\delta_\eta^i(Z)$ values satisfying the condition (11) maintain a positive probability, while the remaining sequences get a zero probability. Note that in the model used in [15], the probability densities of the relevant variables are not updated using information obtained with the arrival of Y_η .

In simple terms, the PV model has three parts representing different aspects of solar power production. The optimal multiplier p_η indicates the level of sunlight on a given day. The change in position of the sun relative to PV panels is described by Y_η , and any deviations from the daily power production due to weather conditions are noted by δ_η . Each of these elements has its own probability representation in the time series, allowing estimation of their values or generation of realizations of them when their exact values are unknown.

C. PV Power Data Sampling

In this section, the process of sampling PV power from the PV power model is described. It is worth noting that the sampling method described in this section is different from the one presented in [15] because of the introduction of multiplicative correction terms δ_η . Another reason for this difference is that the probability densities of the relevant

variables are not updated using [15, eq. (11)]. Generating future PV power samples is needed both for the optimization of the PV amount as detailed in Section V and for the stochastic pump scheduling method presented in Section IV. In the context of PV amount optimization, it is necessary to generate future samples of PV power production for a specified number of PVs. These samples are used to simulate the WDN as if they represented the actual one-year PV power values fueling the entire WDN. On the other hand, for the stochastic pump scheduling method, realizations of PV power production until the day's end, at each instance of pump schedule calculation, are used to estimate the expected value in the cost function [see (22a)]. It is important to note that the samples created for the pump scheduling method are distinct from the one-year samples. This is because, in real-life situations, when we calculate pump schedules, we do not know the future values of these samples, and the same assumption is made in our simulations. The one-year samples are regarded as actual power production values, whereas the samples for the pump scheduling method are projections of these actual values for a one-day horizon before we observe them. The problem of sampling can be stated as follows. Let i_c denote the current time index in day η . Then, the task is to sample realizations of the unseen power values of the current day X_η^i , $i = i_c + 1, i_c + 2, \dots, N_{pv}$. Once a sampling method is identified, obtaining one-year samples involves sequentially generating daily samples for the entire year. For the stochastic pump scheduling method, the sampling method could be used at each instance i_c of pump schedule calculation to generate realizations of PV power generation until the end of the day. The sampling task exhibits a distinction between the period before sunrise and after sunrise due to the availability of power measurements. After sunrise, the power measurements X_η^i , $i = 0, 1, \dots, i_c - 1$ become available at the time i_c , which can be used to update the probability distribution of the variables.

The values of X_η^i are generated by sampling the optimal multiplier p_η and the multiplicative correction terms δ_η^i since Y_η is already calculated at the start of the day η . Before sunrise, p_η values are generated by sampling ϵ_{p_η} from its respective probability distribution and putting the sampled value to the optimal multiplier model (7). The samples of multiplicative correction values δ_η^i are generated by sampling $\zeta_{\delta_\eta}^i$ according to updated distribution (12) and putting those samples into the model of correction terms (10). To sample $\zeta_{\delta_\eta}^i$ values from the updated distributions (12), a rejection sampling method is employed. The $\zeta_{\delta_\eta}^i$ values are generated independently of their normal distributions, and the samples are accepted if the resulting $\delta_\eta^i(Z)$ values, derived by substituting the sampled $\zeta_{\delta_\eta}^i$ values into the correction term model (10), satisfy the following inequality:

$$-\epsilon < \sum_{i=1}^{N_{pv}} Y_\eta^{i2} \delta_\eta^i - \sum_{i=1}^{N_{pv}} Y_\eta^{i2} < \epsilon \quad (13)$$

for some small $\epsilon \in \mathbb{R}_{>0}$. The samples obtained using this method closely approximate the updated distributions (12).

After sunrise, non-zero power production data X_η^i starts coming in, giving more information about the variables δ_η^i and p_η . Let the index corresponding to sunrise be denoted as sr. At time i_c , the δ_η^i , $i = sr + 1, sr + 2, \dots, i_c$ values become deterministic functions of p_η given by the equations $\delta_\eta^i = X_\eta^i / (p_\eta Y_\eta^i)$ for $i = sr + 1, sr + 2, \dots, i_c$ as the values of X_η^i , $i = sr + 1, sr + 2, \dots, i_c$ are available at time i_c . Then, the error values $\zeta_{\delta_\eta}^i = k_i(p_\eta)$ for $i = sr + 1, sr + 2, \dots, i_c$ also become deterministic functions of p_η , where $k_i(p_\eta)$ can be obtained by replacing the correction terms δ_η^i with $X_\eta^i / (p_\eta Y_\eta^i)$ in (10). Given this information, the probability density of p_η is updated using

$$f_{p_\eta | \zeta_{\delta_\eta}^{sr+1}, \dots, \zeta_{\delta_\eta}^{i_c}, \sum_{i=1}^{N_{pv}} Y_\eta^{i2} \delta_\eta^i} \left(x \mid k_{sr+1}(p_\eta), k_{i_c}(p_\eta), \sum_{i=1}^{N_{pv}} Y_\eta^{i2} \right) \quad (14a)$$

$$\propto f_{p_\eta}(x) f_{\zeta_{\delta_\eta}^{sr+1}, \dots, \zeta_{\delta_\eta}^{i_c}, \sum_{i=1}^{N_{pv}} Y_\eta^{i2} \delta_\eta^i | p_\eta} \times \left(k_{sr+1}(p_\eta), \dots, k_{i_c}(p_\eta), \sum_{i=1}^{N_{pv}} Y_\eta^{i2} \mid x \right) \quad (14b)$$

$$= f_{p_\eta}(x) f_{\zeta_{\delta_\eta}^{sr+1}, \dots, \zeta_{\delta_\eta}^{i_c}, \sum_{i=1}^{N_{pv}} Y_\eta^{i2} \delta_\eta^i} \left(k_{sr+1}(x), \dots, k_{i_c}(x), \sum_{i=1}^{N_{pv}} Y_\eta^{i2} \right) \quad (14c)$$

$$= f_{p_\eta}(x) f_{\zeta_{\delta_\eta}^{sr+1}, \dots, \zeta_{\delta_\eta}^{i_c}}(k_{sr+1}(x), \dots, k_{i_c}(x)) \quad (14d)$$

$$f_{\sum_{i=1}^{N_{pv}} Y_\eta^{i2} \delta_\eta^i | \zeta_{\delta_\eta}^{sr+1}, \dots, \zeta_{\delta_\eta}^{i_c}} \left(\sum_{i=1}^{N_{pv}} Y_\eta^{i2} \mid k_{sr+1}(x), \dots, k_{i_c}(x) \right) = f_{p_\eta}(x) f_{\zeta_{\delta_\eta}^{sr+1}}(k_{sr+1}(x)), \dots, f_{\zeta_{\delta_\eta}^{i_c}}(k_{i_c}(x)) \quad (14e)$$

$$f_{\sum_{i=i_c+1}^{N_{pv}} Y_\eta^{i2} \delta_\eta^i} \left(\sum_{i=1}^{N_{pv}} Y_\eta^{i2} - \sum_{i=1}^{i_c} Y_\eta^{i2} X_\eta^i / (x Y_\eta^i) \mid (x Y_\eta^i) \right).$$

The relation (14b) is obtained by applying Bayes' rule $f_{X|Y=y}(x) \propto f_X(x) f_{Y|X=x}(x)$ to (14a). Then, the conditional probability in (14b) is transformed to unconditional probability (14c) by setting $p_\eta = x$. Then, the conditional density rule $f_{X,Y} = f_Y(y) f_{X|Y=y}(x)$ is applied to get (14d). Finally, the independency of $\zeta_{\delta_\eta}^i$ terms are used to separate joint probabilities $f_{\zeta_{\delta_\eta}^{sr+1}, \dots, \zeta_{\delta_\eta}^{i_c}}(\cdot)$ into marginal distributions $f_{\zeta_{\delta_\eta}^i}(\cdot)$ in (14e). The conditional probability $f_{\sum_{i=1}^{N_{pv}} Y_\eta^{i2} \delta_\eta^i | \zeta_{\delta_\eta}^{sr+1}, \dots, \zeta_{\delta_\eta}^{i_c}}$ in (14d) is transformed into the unconditional version in (14e) by setting $\delta_\eta^i = X_\eta^i / (x Y_\eta^i)$ for $i = sr + 1, sr + 2, \dots, i_c$ as the conditions $\zeta_{\delta_\eta}^i = k_i(p_\eta)$ for $i = sr + 1, sr + 2, \dots, i_c$ are equivalent to $\delta_\eta^i = X_\eta^i / (x Y_\eta^i)$.

The summation $\sum_{i=i_c+1}^{N_{pv}} Y_\eta^{i2} \delta_\eta^i$ in (14e) is a linear combination of random variables $Y_\eta^{i2} \delta_\eta^i$ that follow a lognormal distribution as deduced from the model of the correction terms (10). To approximate the probability density of this summation, a method presented in [35] is employed. This method allows for the approximation of the probability density of the sum of two correlated lognormal random variables with a lognormal distribution. Then, the optimal multiplier can be sampled from the updated density (14e). The next task is to sample the corrections terms δ_η^i , $i = i_c + 1, i_c + 2, \dots, sd$ to get the sampled power values X_η^i , $i = i_c + 1, i_c + 2, \dots, sd$, where sd

is the index for sundown. Given the optimal multiplier sample $p_\eta = x$, the probability densities of the $\zeta_{\delta_\eta}^i$ terms are updated as

$$\begin{aligned} & f_{\zeta_{\delta_\eta}^{i_c+1}, \zeta_{\delta_\eta}^{i_c+2}, \dots, \zeta_{\delta_\eta}^{sd} | p_\eta} \left(Z_{\delta_\eta}^{i_c+1}, Z_{\delta_\eta}^{i_c+2}, \dots, Z_{\delta_\eta}^{sd} \middle| x \right) \\ & \propto \begin{cases} f_{\zeta_{\delta_\eta}^{i_c+1}} \left(Z_{\delta_\eta}^{i_c+1} \right), \dots, f_{\zeta_{\delta_\eta}^{sd}} \left(Z_{\delta_\eta}^{sd} \right), & \sum_{i=1}^{N_{pv}} Y_\eta^i \delta_\eta^i = \sum_{i=1}^{N_{pv}} Y_\eta^i \\ 0, & \text{otherwise} \end{cases} \end{aligned} \quad (15)$$

where $\delta_\eta^i = X_\eta^i / (x Y_\eta^i)$ for $i = sr, \dots, i_c$. The $\zeta_{\delta_\eta}^i$ values are sampled using the rejection sampling method from the updated density (15), similar to the way they were sampled before sunrise.

III. WDN MODEL

In this section, a control-oriented model of WDNs is presented. A WDN is composed of various components, such as pipes, demand nodes, pumps, tanks, and more. In this article, WDNs with elevated tanks either directly connected or standing alone are considered. It is assumed that the network has elevated tanks, and as long as the tank water levels are maintained above a certain threshold, the resulting pressure levels in the network will remain within acceptable bounds with the help of valves. Additionally, pumps in the network are considered to be variable-speed pumps, and each pumping station is assumed to have pump inverters, letting the facility control the flow rate of the pumps. The hydraulic head is a measure of fluid pressure, representing the height of a static fluid column at a specific point. The loss of hydraulic head in a pipe can be estimated using the Hazen-Williams equation as

$$\Delta h = h_1 - h_2 = Kq|q|^{0.852} \quad (16)$$

where K is the pipe resistance, which depends on the physical features of a pipe, such as the diameter and length of the pipe, q is the flow rate, and h_1 and h_2 are the heads at the two ends of the pipe. At each node j , the inflow of water is equivalent to the outflow of water, which can be expressed as

$$\sum_{i \in \mathcal{N}_j} q_{ij} = d_j \quad (17)$$

where q_{ij} is the flow entering the node j from node i and d_j is the demand at node j , which is the water requested by the user at node j . The symbol \mathcal{N}_j denotes the set of neighbor nodes of node j . Note that q_{ij} is positive if the flow is from node i to the neighbor node j and negative, vice versa.

Valves are used to control the flow or the pressure at specific points in a WDN. There are various types of valves such as pressure-reducing valves (PRVs) and flow control valves (FCVs). PRVs limit the pressure at a point in the WDN. FCVs limit the flow below a certain level. In this work, it is assumed that the valves in the network are controlled according to pre-defined strategies to manage the pressure and flow at specific locations.

The water tanks are the only dynamic components within a WDN, and their behavior can be described as

$$A_j \dot{h}_j = \sum_{i \in \mathcal{N}_j} q_{ij} \quad (18)$$

where A_j represents the cross-sectional area of the tank identified by index j , h_j is the tank level, q_{ij} is the flow entering the tank, and \mathcal{N}_j denotes the set of neighbor nodes of the tank j . Tank levels change according to the flow passing through the pipes connected to the tanks. These flow rates q_{ij} are affected by the valve control methods and head loss equations (16) of the pipes in the network, and mass balance equations (17) of the demand nodes in the whole network. The flow rates of the pipes connected to the tanks q_{ij} can be written as functions of demand at each node, tank levels, and the amount of water coming from the pumps as

$$q_{ij} = f_{ij}(h, q, d) \quad (19)$$

where $h \in \mathbb{R}^n$, $q \in \mathbb{R}^m$ are the vectors containing the tank levels and pump flow rates, respectively, and $d = [d_1, d_2, \dots]^T$ is the vector containing the demands of all the nodes. Combining (19) with the tank level equations (18) gives

$$A_j \dot{h}_j = \sum_{i \in \mathcal{N}_j} f_{ij}(h, q, d). \quad (20)$$

Note that the functions f_{ij} are nonlinear because of the nonlinearity in the head loss equations (16) of the pipes and the potential nonlinearity introduced by the valves. Since not all facilities keep track of water demands at each node, it is assumed that such detailed data is unavailable. Instead, it is assumed that the total demand of the area served by the pumps can be estimated based on historical data. Consequently, the tank dynamics can only be approximated because of the lack of knowledge of individual demands. To approximate the tank dynamics described by (20), a linear model is used and written as

$$\dot{h}(t) = Ah(t) + B_1 u(t) + B_2 d_a(t) \quad (21)$$

where $h(t) \in \mathbb{R}^n$ represents the tank levels measured in meters, $A \in \mathbb{R}^{n \times n}$, $B_1 \in \mathbb{R}^{n \times m}$, and $B_2 \in \mathbb{R}^{n \times 1}$ are constant system matrices, $d_a(t)$ is the aggregated demand of the area supplied by the pumping stations at time t measured in m^3/s , and $u(t) \in \mathbb{R}^m$ is the input containing the flow rate of the water exiting the pumping stations measured in m^3/s . A linear model is chosen for control-oriented modeling because it reasonably approximates the tank level evolution given by the EPANET model, and it is computationally efficient when used in optimization-based control methods [see (22)]. A more complex model can be built, and it would be more accurate than the linear model (21). However, with a complex model, the computation of the control input would last longer if the model were to be used with an optimization-based method. Therefore, the use of such a complex model to compute the control input in an online fashion may not have been efficient. Note that the fast pipe dynamics in the network are not accounted for in our model. This is because tank dynamics are so slow that the effect of the fast pipe dynamics on the tank levels is negligible. The matrices A , B_1 and B_2 are obtained by applying least squares to fit linear functions of tank levels, pump flows, and the aggregated demand to the flow rates of the pipes q_{ij} connected to the pumps, using the data generated by the EPANET model of the WDN.

IV. STOCHASTIC ECONOMIC MODEL PREDICTIVE CONTROLLER FOR PUMP SCHEDULING

This section presents a predictive control-based pump scheduling method for WDNs that are powered by grid-connected PV panels. The approach, initially proposed in [15], focuses on minimizing the economic costs involved in WDN operation while satisfying network constraints. To implement this method, a central computer and a communication infrastructure linking tanks and pumps to this central unit are assumed prerequisites. Consequently, the central computer has access to tank levels and the ability to solve the control problem. Upon resolving the problem, it communicates the calculated flow rates to the pumping stations, where local controllers subsequently modify the pump flow rates accordingly. The problem at time t is formulated as

$$\min_{u'_0, u'_1, \dots, u'_{N(t)-1}} \sum_{j=0}^{N(t)-1} \mathbb{E}[J(h'_j, u'_j, t, P_{pv}, c)] \quad (22a)$$

$$h'_j = A_d h'_{j-1} + B_d u'_{j-1} + B_{d2} d'_a(j-1) \quad (22b)$$

$$h'_0 = h(t) \quad (22c)$$

$$u'_j \in \mathcal{U} \subseteq \mathbb{R}^m \quad (22d)$$

$$h'_j \in \mathcal{H} \subseteq \mathbb{R}^n \quad (22e)$$

$$h'_{N(t)} \in \mathcal{H}_{tf} \subseteq \mathbb{R}^n \quad (22f)$$

where $\mathbb{E}[J(h'_j, u'_j, t, P_{pv}, c)]$ is the expected value of economic cost function $J(h'_j, u'_j, t, P_{pv}, c)$, $h' = [h'_1, \dots, h'_{N(t)}] \in \mathbb{R}^{n \times N(t)}$ is the predicted future states, $u'_j \in \mathbb{R}^m$ is the input vector, d'_a is the estimated aggregated demand of the area supplied by the pumping stations, P_{pv} is the PV power, c is the electricity prices in €/kW, $N(t)$ is the prediction horizon, $\mathcal{U} \subseteq \mathbb{R}^m$ and $\mathcal{H} \subseteq \mathbb{R}^n$ denote the input and state constraints, respectively, and $\mathcal{H}_{tf} \subseteq \mathbb{R}^n$ is the terminal state set. Equation (22b) represents the discretized version of the continuous system (21). At each time step, which is separated by a time interval of Δ_t , the optimization problem (22) is solved and the first term u'_0 of the optimal input sequence $\mathbf{u}' = [u'_0, \dots, u'_{N(t)-1}] \in \mathbb{R}^{m \times N(t)}$ is applied to the system.

The minimum and maximum flow rate capacity of the pumps determines the input constraints for the system. These conditions are expressed as

$$\mathcal{U} = \{[u_1, \dots, u_m]^T \in \mathbb{R}^m | \forall i : 0 \leq u_i \leq \bar{u}_i\} \quad (23)$$

where $\bar{u}_1, \dots, \bar{u}_m$ are upper flow limits. The constraints on tank levels are established by the tank capacities and the necessary water reserve required at all times to handle emergencies effectively. The set \mathcal{H} is defined as

$$\mathcal{H} = \{[h_1, \dots, h_n]^T \in \mathbb{R}^n | \forall i : \bar{h}_i \leq h_i \leq \bar{h}_i\}. \quad (24)$$

In this work, it is assumed that the pressures in the WDN would be regulated by the elevated tanks and PRVs. Because of this, pressure constraints are not included in the controller formulation (22). Water losses because of the pressure levels in the network are assumed to be captured by the total aggregated demand d_a .

Linear model (22b) provides only an approximation of the WDN and does not capture its complete representation. Also,

it is assumed that the future values of the total demand d_a can only be estimated. Consequently, the true state trajectories h will differ from the predicted states h' . Although the predicted states h' adhere to the state constraints by the construction of the controller problem (22), the actual trajectories h may not satisfy these constraints, particularly when the predicted states approach the boundaries of the state constraint set \mathcal{H} . Even small deviations of actual trajectories h from the predicted states h' can result in violations of the state constraints when the predicted states are close to the boundaries. To address this issue, exponential barrier-like functions are used to keep the predicted trajectories away from the boundaries of state constraints so that actual trajectories also satisfy the state constraints. First, state constraints (24) are rewritten as

$$F_i(h) \leq 0, \quad i = 1, 2, \dots, 2n \quad (25)$$

where $F_{2i}(h) = \bar{h}_i - h_i$, $i = 1, \dots, n$ and $F_{2i-1}(h) = h_i - \bar{h}_i$, $i = 1, \dots, n$. The cost function terms are then defined as

$$J_{hi}(h) = e^{a_i(F_i(h)+b_i)} \quad i = 1, \dots, 2n \quad (26)$$

where $a_i, b_i \in \mathbb{R}_{>0}$. The parameters a_i and b_i specify a region next to the boundaries of the state constraint set \mathcal{H} , wherein the cost function J_{hi} obtains high values. When an optimizer solves the problem (22), it automatically finds solutions h' that do not enter this expensive region. As long as the predicted states h' stay away from the edges of \mathcal{H} , the actual states $h(t)$ will also stay within those limits, provided that the deviation of the actual states from the predicted remains sufficiently small. In simple terms, the cost function terms J_{hi} penalize being too close to the edges of \mathcal{H} . Therefore, predicted states h' stay away from the edges, ensuring that the actual paths $h(t)$ stay within \mathcal{H} .

The cost function $J(h'_j, u'_j, t, P_{pv}, c)$ incorporates the expenses associated with purchasing electricity from the electrical grid. The primary energy source for the WDN is obtained from PV panels, and in cases where the power generated by the PV panels is less than the power required by the WDN, additional electricity is purchased from the grid to meet the demand. When PV power is more than the power required by the WDN, no electricity is bought from the grid. Therefore, the electricity purchased from the grid is represented as

$$P_{\text{grid}}(t) = \max(0, P_p(t) - P_{pv}(t)) \quad (27)$$

where P_{grid} is the power purchased from the grid, P_p is the total power used by the pumps, and $P_{pv}(t) = X_{\eta}^i$ is the power generated by the PV panels, where η and i refer to the specific day and time index within that day that aligns with the moment t . Note that the right-hand-side of (27) is equal to the rectified linear unit (RELU) function of $P_p - P_{pv}$. As the grid power P_{grid} is a non-differentiable function, RELU, of $P_p - P_{pv}$, it is approximated with the softplus function

$$\text{sp}(P_p(t) - P_{pv}(t)) = \frac{1}{\beta} \log(1 + \exp(\beta(P_p(t) - P_{pv}(t)))) \quad (28)$$

where β is a constant. This function is widely used as a smooth approximator of the RELU function. The power consumed by the pumping station i is equal to $u_i(p_i^{\text{out}} - p_i^{\text{in}})/\lambda_i$, where u_i is the flow out of the pumping station, p_i^{out} and p_i^{in} are the outlet and

inlet pressures of the station measured in N/m^2 , and λ_i is the efficiency. It is assumed that each pumping station has several pumps of the same size connected in parallel. The number of active pumps is controlled by a low-level controller to minimize power consumption. Then, the efficiency of the pumping stations remains relatively stable, allowing the assumption of constant efficiency, $\lambda_i = \lambda, \forall i \in \{1, 2, \dots, m\}$ [36]. The outlet pressures $p^{\text{out}} \in \mathbb{R}^m$ are given as the output of the linear model

$$p^{\text{out}}(t) = Ch(t) + Du(t) \quad (29)$$

where $C \in \mathbb{R}^{m \times n}$ and $D \in \mathbb{R}^{m \times m}$ are found using system identification on data generated by the EPANET model. The inlet pressures $p^{\text{in}} \in \mathbb{R}^m$ refer to the pressures of distinct water reservoirs where pumps draw water. It is assumed that these pressures remain constant, implying that the water levels in the associated reservoirs are considered to be unchanging. While daily variations in reservoir water levels are possible, this assumption is made due to the lack of available data, as the reservoir levels on EPANET models are always constant. However, in real situations, inlet pressures can be modeled more accurately like outlet pressures are modeled in (29) to capture daily variations if the measurements for water reservoir levels are available. The power used by the pumps at time t is then equal to $P_p(t) = u(t)^T(p^{\text{out}}(t) - p^{\text{in}}(t))/\lambda$. The overall cost function includes both the electricity expense term and the constraint barrier functions

$$J(h(t), u(t), t, P_{\text{pv}}, c) = \sum_{i=1}^{2n} J_{h_i}(h(t)) + \sum_{k=0}^{\frac{\Delta_t}{\Delta_{\text{pv}}} - 1} c(t) \text{sp}(P_p(t) - P_{\text{pv}}(t + k\Delta_{\text{pv}})). \quad (30)$$

It is assumed that the sampling time for PV data, Δ_{pv} , is shorter than the system sampling time, Δ_t , resulting in multiple PV power values being available within the time period from t to $t + \Delta_t$. The electrical costs are then evaluated individually for each distinct PV power value $P_{\text{pv}}(t + i\Delta_{\text{pv}})$ in the interval.

The control problem given in (22) is solved with a scenario-based approach in which $S \in \mathbb{N}$ number of scenarios are generated. The only stochastic variable appearing in the cost function $J(h(t), u(t), t, P_{\text{pv}}, c)$ is the PV power P_{pv} . The PV power values P_{pv} are sampled following the procedure outlined in Section II-C. However, there is a difference in the approach for sampling the ζ_{δ_η} values which are obtained from distributions (12) before sunrise and (15) after sunrise. In Section II-C, a rejection sampling technique is used so that (13) is satisfied. Here, the rejection sampling technique is not used, and all sets of ζ_{δ_η} generated from independent normal distributions are accepted. This adjustment is made to prioritize faster sampling, even though it results in a decrease in the level of approximation to the actual probability distribution. Consequently, the samples produced using this method may not be as precise as those generated following the approach outlined in Section II-C. This discrepancy leads to a less accurate approximation of the cost function defined in (22). As a result, the pump schedules found through approximating the solution of (22) with this sampling method are suboptimal, implying diminished cost-saving compared to the case where sampling is done as explained in Section II-C. If this control

method were to be implemented for a real WDN without speed concerns, the rejection sampling could still be used. However, in this article, the simulations for the optimization of PV values take a significant amount of time (see Section V). Therefore, to accelerate the simulations, the rejection sampling was not used in this article.

The daily water demand profiles are closely similar, with only small variations. This is used to increase the chance of finding a solution to the optimization problem (22). The idea revolves around maintaining the tank levels at the start of each day within a specific range. If a pumping schedule can steer the system's state trajectories close to their initial values by the end of a day, it is probable that a similar schedule could be applied on the next day since the tank levels at the beginning of each day are close to one another. Therefore, the horizon $N(t)$ is chosen to target the beginning of each day and expressed as

$$N(t) = (T_{\text{day}} - t \bmod T_{\text{day}})/\Delta_t \quad (31)$$

where T_{day} is the duration of a whole day. To define the terminal set \mathcal{H}_{tf} in controller problem (22) that the trajectories should return to at the start of each day, the optimal periodic trajectory of the system should be defined first. The optimal periodic trajectory is defined just as in [37] as the solution of

$$(\mathbf{u}^*, \mathbf{h}^*) = \arg \min_{u_i, h_i} \sum_{i=0}^{(T_{\text{day}}/\Delta_t)-1} J(h_i, u_i, t, P_{\text{pv}}^*, c^*) \quad (32a)$$

$$h_i = A_d h_{i-1} + B_{d1} u_{i-1} + B_{d2} d_a^*(i-1) \quad (32b)$$

$$u_i \in \mathcal{U} \subseteq \mathbb{R}^m \quad (32c)$$

$$h_i \in \mathcal{H} \ominus \mathcal{W} \subseteq \mathbb{R}^n \quad (32d)$$

$$h_0 = h_{T_{\text{day}}/\Delta_t} \quad (32e)$$

where $d_a^*(i)$ $i = 0, 1, \dots, (T_{\text{day}}/\Delta_t) - 1$ is the average daily demand profile obtained from the past measurements, P_{pv}^* is the average daily PV profile, c^* is the average daily electricity price profile, and the set \mathcal{W} is a design choice. The resulting state trajectory $\mathbf{h}^* = [h_0^*, \dots, h_{T_{\text{day}}/\Delta_t}^*] \in \mathbb{R}^{n \times (T_{\text{day}}/\Delta_t + 1)}$ is the optimal periodic trajectory as dictated by the periodicity constraint (32e). The terminal set for the controller problem (22) is chosen as the points close to the $h_{T_{\text{day}}/\Delta_t}^*$. This decision is based on the idea that if the actual tank levels would revolve around the optimal periodic trajectory \mathbf{h}^* , the associated economic costs would remain low. By driving tank levels close to $h_{T_{\text{day}}/\Delta_t}^*$ at the end of each day, the actual tank levels would remain close the optimal periodic trajectory \mathbf{h}^* , whenever the power, electricity price and demand profiles of a day are close to the average profiles $P_{\text{pv}}^*, c^*, d_a^*$. After finding the solution to (32), the terminal state constraint set \mathcal{H}_{tf} for the controller problem (22) then can be written as

$$\mathcal{H}_{tf} = \mathcal{B}_r(h_{T_{\text{day}}/\Delta_t}^*) \quad (33)$$

where $\mathcal{B}_r(h_{T_{\text{day}}/\Delta_t}^*)$ is the open ball centered at $h_{T_{\text{day}}/\Delta_t}^*$ with radius r . If the problem (22) becomes infeasible at any time step t , the second term of the input sequence from the previous step $u_1^{t-\Delta_t}$ is applied to the network. The reason behind using the narrower set $\mathcal{H} \ominus \mathcal{W}$ as the state constraint set, rather than \mathcal{H} , when computing the optimal periodic trajectory as

defined in (32), is to prevent the violation of state constraints by the state trajectories $h(t)$ when problem (22) becomes infeasible. Normally, if the optimization problem is feasible, the state trajectories $h(t)$ are expected to stay close to the optimal periodic trajectory \mathbf{h}^* because of the terminal state constraint \mathcal{H}_{tf} . However, the state trajectories $h(t)$ might deviate more from the optimal periodic trajectory \mathbf{h}^* when the optimization problem (22) becomes infeasible. To mitigate this, by confining the optimal periodic trajectory within the narrower set $\mathcal{H} \ominus \mathcal{W}$ instead of \mathcal{H} , it becomes more likely that the state trajectories $h(t)$ will remain within the bounds of the state constraint set \mathcal{H} even when the optimization problem is infeasible. The design choice for set \mathcal{W} involves a trade-off between feasibility and cost-effectiveness. If it is chosen to be big, the optimal periodic trajectory \mathbf{h}^* will be confined to the more inner parts of \mathcal{H} , which will lead to state trajectories $h(t)$ revolving more in the inner side of \mathcal{H} , resulting in a less cost-effective but safer setting in terms of satisfaction state constraints. Conversely, a smaller \mathcal{W} has the opposite effect, less conservative and more cost-effective.

In this work, the error set \mathcal{W} is chosen as the set containing the possible one-step errors between the linear model (22b) and the actual EPANET model of WDN. This set is found by investigating the dataset generated by the EPANET model and taking the difference between the tank levels generated by EPANET and the linear model (see Section VI-A).

V. OPTIMIZING PV INSTALLATION AMOUNT

This section presents a new method to optimize the amount of PV installation for a WDN. The objective is to find the amount of PV that minimizes the overall expected cost of the WDN throughout the lifespan of the PV panels, considering both operational expenses (OPEXs) and capital expenditures (CAPEXs). OPEX refers to the costs associated with maintaining the PV panels and purchasing energy from the grid to power the pumps, while CAPEX encompasses the installation costs of the PV panels. The problem is written as

$$\min_x \mathbb{E} \left[J_c(x) + \sum_{i=1}^{\ell_{pv}} J_o(x, P_{\text{grid}}^i) \right] \quad (34)$$

where x is the PV amount measured by its initial power production under normal conditions $x = P_{\text{STC}}^1$ with P_{STC}^i representing the i th year's power production under normal conditions, P_{grid}^i representing the power bought from the grid for the year i , $J_c(x)$ is the CAPEX, $J_o(x, P_{\text{grid}}^i)$ is the yearly OPEX, and ℓ_{pv} is the lifespan of the PVs. The expected cost is determined under some assumptions. First, it assumes that the climate will remain constant throughout the lifespan of the PV panels. Consequently, the PV power model, developed using historical weather data, provides an accurate representation of future power production. Although this assumption is not entirely accurate, the potential decrease in PV power production is estimated to be small in the lifespan of a PV panel, at most 4% in 25 years according to Fig. 3 in [38]. Second, it assumes that the water demand profile will remain unchanged throughout the lifespan of the PV system. This assumption is also not entirely true as the population



Fig. 3. WDN of Randers. The pumping stations to be controlled are shown in red and the remaining in blue. Tanks are shown with a “T” shaped symbol in yellow.

in the area supplied by the pumps might change because of various reasons. Let us consider a situation where a facility determines an approximate optimal amount of PVs with this assumption and installs them. If the population in the serviced area grows, a reasonable expectation would be an increased optimal number of PVs. In this case, the initial installation of PVs does not hinder achieving the new optimal amount, as the WDN facility can add more PVs to match the increased demand. Conversely, if the population decreases for unknown reasons, the anticipated optimal PV amount for the reduced population would be smaller than the initial installation. In such a case, it could be argued that the initially installed PVs were excessive. However, due to the difficulty in predicting these demographic shifts, the assumption of constant water demand is retained. Additionally, the linear model (22b) is assumed to accurately represent the WDN throughout the lifespan of PVs. This means that factors such as the aging of WDN elements, including pipes, are not considered in this work. This assumption is made because of the challenge of anticipating alterations in the WDN and their potential impact on the parameters of the linear model.

Various factors, such as the panel type, market competition within the PV industry, the purchasing country, and the scale of the installation can influence the pricing of PV panel installation. Typically, as the quantity of PV panels increases, the cost per kilowatt (kW) tends to decrease. While this article adopts a constant unit cost based on an average price, a more detailed model could be used by engaging in discussions with a specialized PV installation company to consider tailored unit costs. Using a more detailed cost model for installation could lead to more savings in the total costs of the WDN. This is because, usually, as you increase the number of PV panels, the cost per panel goes down, thanks to discounts from buying in bulk. The installation cost can be written as

$$J_c(x) = a_{\text{ins}} x \quad (35)$$

where $a_{\text{ins}} \in \mathbb{R}$ is the constant unit cost. OPEX covers the maintenance costs and the costs of purchasing power from the grid and can be written as

$$J_o(x, P_{\text{grid}}^i) = a_m x + \sum_{j=1}^{365} \sum_{l=1}^{T_{\text{day}}/\Delta_{pv}} c(t) P_{\text{grid}}^i(t) \Delta_{pv} \quad (36)$$

where $a_m \in \mathbb{R}$ is the unit maintenance cost, $t = jT_{\text{day}} + l\Delta_{\text{pv}}$ is the time of the year, and $P_{\text{grid}}^i(t)$ is the power bought from the grid at time t in year i . The power bought from grid $P_{\text{grid}}^i(t)$ depends deterministically on the yearly PV power samples $P_{\text{pv}}^i(t)$. The statistical properties of $P_{\text{pv}}^i(t)$ vary with the year index i due to changes in the standard test condition power P_{STC}^i over the years. The standard test condition power $P_{\text{STC}}^i = \lambda_{\text{pv}}^i x$, $i = 1, \dots, \ell_{\text{pv}}$ changes over time because the efficiency of the solar panels, λ_{pv}^i , decreases as they age. To simplify the analysis, a constant efficiency model for the PVs is used. This ensures that the standard test condition power $P_{\text{STC}} = \lambda_{\text{pv}} x$ remains constant throughout the lifespan, with $\lambda_{\text{pv}} \in \mathbb{R}$ representing the constant efficiency whose value is set to the average efficiency over the lifespan of PVs $\lambda_{\text{pv}} = (1/\ell_{\text{pv}}) \sum_{i=1}^{\ell_{\text{pv}}} \lambda_{\text{pv}}^i$. By assuming this, the parameters of the PV power model can be kept constant over the years, ensuring that the statistical properties of the yearly solar power production $P_{\text{pv}}^i(t)$ and the power bought from the grid $P_{\text{grid}}^i(t)$ remain the same over the years. Therefore, the expected values $\mathbb{E}[J_o(x, P_{\text{grid}}^i)]$ and $\mathbb{E}[J_o(x, P_{\text{grid}}^j)]$ are equal for any years $i_1, i_2 \in \mathbb{N}$. Now, the dependency of the OPEX on the year index can be removed by defining $J_o(x) = J_o(x, P_{\text{grid}}^{i_1})$ for some fixed year $i_1 \in \mathbb{N}$. This simplification enables reformulation of the problem (34) as

$$\min_x \mathbb{E}[J_c(x) + \ell_{\text{pv}} J_o(x)]. \quad (37)$$

The reasoning behind these assumptions and the removal of the OPEX dependency on the year index is to avoid extensive computation time required to evaluate the term $\sum_{i=1}^{\ell_{\text{pv}}} J_o(x, P_{\text{grid}}^i)$ in (34) for a given value of x . Such calculations would significantly prolong the optimization procedure. We have tried using neural networks to approximate the behavior of the predictive controller, which would have expedited the cost calculation. This approach would have minimized the time needed to simulate the network over the lifespan of PVs and calculate the term $\sum_{i=1}^{\ell_{\text{pv}}} J_o(x, P_{\text{grid}}^i)$. However, the accuracy of the neural network approximations of the controller was insufficient. As a result, it is decided to maintain the assumption of a constant efficiency λ_{pv} .

To approximate the total cost $\mathbb{E}[J_c(x) + \ell_{\text{pv}} J_o(x)]$ for a given amount of PVs x , first step is generating the synthetic historical PV data for the given value $P_{\text{STC}} = \lambda_{\text{pv}} x$ and using the physical production model (1). Note that the same historical weather measurements are used for each x . Then, the parameters of the time series representations (7), (8), and (10) of the PV production model defined by (4)–(6) are fit to the synthetic historical PV data. Then, one year of PV production data is sampled using the PV model. Subsequently, the system's tank levels are simulated for one year by treating the sampled one-year PV power data as the real PV production data. The pump flow rates are derived from solving the control problem (22) at each time step. For this, the terminal condition of the control problem (22) is determined by solving (32e). At each step, the amount of power purchased from the grid $P_{\text{grid}}(t)$ is calculated using (27) after the pump flow rates are found. For the simulation of WDN and evolving the tank levels, the approximate system model (22b) is used instead of the actual

EPANET model of the network, as the latter is computationally intensive. Despite using the approximate model, it still takes approximately 6 h to compute the yearly OPEX $J_o(x)$ for a given x and the yearly power profile on a computer with Intel i5-1135G7 2.40 GHz processor and 32 GB RAM. Due to time constraints, a single sample of the yearly PV profile is used to approximate $\mathbb{E}[J_c(x) + \ell_{\text{pv}} J_o(x)]$ for a given x . While acknowledging the limitation of having only one sample, it is expected to provide a reasonable approximation of the expected value. This is because essentially 365 days are sampled, potentially covering both days when the generated solar power is more than the seasonal average and days when the generated solar power is less than the seasonal average in a year. A similar argument can be made about the intraday variations in PV power production. Using the multiplicative correction term model (10), approximately 365×12 h of intraday variations are generated for a year, providing a reasonable representation of various intraday scenarios.

The closed form of the function in (37) is not known as the calculation of the cost involves solving the optimization problem (22), which also means gradient information for the cost function is also not available. Therefore, a direct search method is used, specifically the Nelder-Mead method [39], to iteratively update PV amounts and find the amount minimizing the cost function. The Nelder-Mead method is particularly suitable for this problem, as it shows significant improvements in the initial steps and requires fewer evaluations of the cost function compared to other methods [40]. This is important for this application as the calculation of the cost function (37) requires the simulation of the model (22b) for a year, which is extremely time-consuming.

The sampling of one-year PV data is done in a sequential process. Initially, the daily profile of the first day is assigned the value from one year ago, $Y_\eta = Y_{\eta-365}$, where $Y_{\eta-365}$ is obtained from the synthetic power data. Then, the optimal multiplier p_η and the multiplicative correction values δ_η for the current day η are sampled using, respectively, (7) and (8), and the power data X_η is obtained. Following that, the daily power profile is updated $Y_{\eta+1}$, and the process continues to the next day. The sampling of the optimal multiplier p_η and the multiplicative correction values δ_η are done as explained in Section II-C. The overall process of calculating the yearly OPEX $J_o(x)$ is given in Algorithm 1.

VI. APPLICATION

The presented method is used to find an approximation of the optimal installed PV amount for the WDN of Randers, a Danish city. The EPANET model of the WDN, illustrated in Fig. 3, is used to get the relevant information about the network. The network contains 4549 nodes and 4905 links connecting them and includes eight pumping stations. Out of these, six are shown in the figure while the remaining two are located where tanks are placed. Our goal is to control the flow rates of the two pumping stations, shown in red color in Fig. 3, and find an approximation of the optimal PV amount for them, while the others operate according to pre-determined strategies. The decision to control and determine PV amounts for only two stations was made in collaboration with the

Algorithm 1 Calculation of Yearly OPEX $J_o(x)$

Input $x, Y_{1-365} \triangleright$ PV Amount and the daily Power profile for the first day

Output $J_o(x)$

procedure

$P_{STC} \leftarrow \lambda_{pv} x$

Produce synthetic PV data with historical weather measurements, namely irradiance (G), ambient temperature (T_{amb}), and wind speed (V_{wind}), using the physical model equations (1), (2), (3).

Fit the parameters of the time series representations (7), (8), (10) of the PV production model defined by equations (4), (5), (6) to the synthetic historical PV data.

$Y_1 \leftarrow Y_{1-365}$

for $k \leftarrow 1$ **to** 365 **do**

 Sample p_k using (7), (8)

 Sample $\delta_k^i, i = 1 \dots T_{day}/\Delta_{pv}$ using (7), (8)

$X_k \leftarrow p_k Y_k \odot \delta_k$

 Get Y_{k+1} using (5)

Get the terminal state set using (32).

for $i \leftarrow 1$ **to** $365 * 24$ **do**

 Solve the problem (22) and find the input $u(i\Delta_t)$.

 Calculate $P_{grid}(i\Delta_t)$ with (27)

 Simulate the system with (22b).

Calculate $J_o(x)$ using (36).

WDN company due to limited space in other pumping stations for PV installations. Also, there are limited opportunities to optimize energy usage in the other pumping stations since they mainly serve areas without tanks, making significant flow rate adjustments difficult due to pressure requirements. The same method could be used for any number of pumping stations in the network. The only difference would be the identification of the parameters in the linear model (21) for the desired pumping stations to be controlled. The stations that will be controlled provide water mostly to the high zone (HZ) and low zone (LZ), as shown in Fig. 3. Additionally, there are three tanks in the HZ, two of them connected with a pipe and the third standing alone.

In the linear model (21), the number of inputs is equal to $m = 2$ since two pumping stations are being controlled. The number of state variables is equal to the number of tanks in the network. However, two of the tanks are directly connected by a large pipe. This means that whenever there is a difference in their water levels, water quickly moves through the pipe connecting them. As a result, the water levels in both tanks, h_1 and h_2 , are almost always very close $h_1 \approx h_2$. Therefore, we can treat the levels of these two tanks, h_1 and h_2 , as a single state variable $h_{1,2}$. To identify the values of constant matrices A, B_1, B_2 in the model (21), first the EPANET model is simulated for 20 days with a 5-min simulation time step, generating data for around 6000 time instances. For each day, different values for the initial tank levels are used, and flow rates of the two controlled pumping stations are set to values varying from 0 to maximum flow rate u_{\max} . Control laws for the remaining pumping stations are predefined within the

EPANET model. Next, the flow of the pipes connected to the tanks is approximated as linear functions of the tank levels $h(t) = [h_{1,2}(t) \ h_3(t)] \in \mathbb{R}^2$, pump flows $u(t) = [u_1(t) \ u_2(t)] \in \mathbb{R}^2$, and the aggregated demand $d_a(t)$ using the simulation data. The total demand for the HZ and LZ is used as the aggregated demand in the model since these areas are primarily supplied by the controlled pumps. The linear model of the system (21) is obtained by putting the linear approximations of the pipe flows into the tank level equation (18). Note that since other pumping stations were controlled using predefined laws in EPANET, their effect on the evolution of the tank levels is reflected in the data generated by EPANET. Therefore, the network model (18) contains information about the effect of other pumping stations as the parameters of the model are fit using this dataset. Also, the effect of the other pumping stations on the levels of the three tanks in HZ is small, either because the capacities of pumping stations are small or they mostly supply water to the other areas.

A. Simulation Results

This section presents the simulation setup and results for the Randers WDN. The historical weather measurement data is taken from the Open Data API of the Danish Meteorological Institute (DMI). The values of the parameters $k_i, i = 1, \dots, 6$ are given in [16], and the values of μ_0, μ_1 in [41] are used for the physical model (1). From the data generated by EPANET, we empirically identified the set of possible differences between the one-step tank level change estimation of the linear model (22b) and the corresponding change in the EPANET model $h_1^t - h(t + \Delta_t)$. This range was determined to be $[-0.1, 0.1] \times [-0.1, 0.1]$. The parameters of the exponential barrier functions J_{hi} were then set to $a_i = 80$ and $b_i = 0.2$ for all i . With these parameter choices, the values of J_{hi} become significantly higher than the other terms in the cost function (30) within the region where $F_i(h) \in [-0.1, 0] \times [-0.1, 0]$, as defined in (25). Therefore, as long as the next predicted state h_1^t is not in the region where J_{hi} takes high values, the actual next state $h(t + \Delta_t)$ will satisfy the state constraints $F_i(h) \leq 0$. This is because $h_1^t - h(t + \Delta_t)$ is found to be always smaller than 0.1 in our EPANET data analysis. In the formulation of optimal periodic trajectory, \mathcal{W} is also defined as the set containing all possible one-step errors $h_1^t - h(t + \Delta_t)$, that is, $\mathcal{W} = [-0.1, 0.1] \times [-0.1, 0.1]$. Both decisions regarding a_i, b_i parameters and \mathcal{W} are a matter of design. Opting for a larger \mathcal{W} could be a safer choice in terms of state constraint satisfaction but this might make the control less cost-effective. The same could be said about a_i, b_i parameters and the region defined by them. The maximum tank levels are 3 m for h_1 and h_2 , and 2.8 m for h_3 , while the minimum tank level is set to half-full. The maximum pump flow is set to 100 L/s and the sampling time Δ_t to 1 h, so the control input is recalculated at each hour. In all simulations, the yearly electricity prices $c(t)$ are set to the day-ahead electricity prices in Denmark from the year 2022. The daily total demand of HZ and LZ $d_a(t), t = 0, \Delta_t, \dots, T_{day} - \Delta_t$ is obtained from the EPANET model, and the network model (22b) is simulated using these values in all simulations. Given the assumption of constant daily demand, it is inferred that water losses, encapsulated in

the total demand, remain stable throughout the entire lifespan of PVs. We can only use an estimation d_a^t of the daily total demand values in our calculations of pump schedules (22) since we assume that the total demand can only be estimated. However, historical water demand data of Randers WDN is not available, so the total demand $d_a(t)$ cannot be estimated. Instead, a historical demand data set from another facility in Denmark is used. First, the daily average demand of the demands in the dataset is found. Then, a small perturbation signal is obtained by taking the difference between the daily average demand and the demand on a specific day. This small perturbation is added to the actual total demand $d_a(t)$ $t = 0, \Delta t, \dots, T_{\text{day}} - \Delta t$ of Randers WDN resulting in a perturbed daily demand that imitates the estimated demand d_a^t . In each simulation, a different day from the data set is used, so the assumed estimated demand is different each time. The calculations are based on the assumption that the PV lifespan is 25 years $\ell_{\text{pv}} = 25$ [42]. However, we also report the cost values for different lifespan values. In [43], it is reported that the degradation rate of the PVs in Aalborg, Denmark is around 0.15%. The value of average efficiency λ_{pv} is set using this information to its average value over the lifespan of the PVs as $\lambda_{\text{pv}} = 1 - 0.0015\ell_{\text{pv}}/2$. The unit maintenance cost for PVs is set to $a_m = 17\text{€}/\text{kW}$ [44]. The unit installation price is set to $a_{\text{ins}} = 2\text{€}/\text{W}$, which is decided based on the market prices in Denmark.

Before presenting the results of the PV sizing algorithm, we first discuss the accuracy of the PV model presented in Section II. There are two main differences in the PV model compared to the one in [15]: the inclusion of a seasonal component in the optimal multiplier model (7), and the usage of the multiplicative correction terms δ_η in the PV model (4) instead of the additive ones in [15]. The seasonal component is included to capture datasets with seasonal characteristics, as illustrated by the synthetic dataset's seasonality shown in Fig. 2. An ARMA model, as used in [15], would not effectively represent such seasonal patterns. The performance of multiplicative correction terms was assessed through experiments on synthetic PV power data. The models were trained on a portion of the dataset and evaluated on the test set, where the multiplicative correction terms achieved significantly higher log-likelihood values (-4.8729×10^3) compared to the additive correction terms (-1.4453×10^5), demonstrating their ability to better capture the variability and statistical properties of PV power production.

The overall cost of the WDN is determined by simulating the linear model (21). The complete calculation process for obtaining the total cost associated with a specific number of PVs typically requires approximately 6 h. The calculated cost values for different amounts of PVs encountered throughout the iterations of the Nelder-Mead method are reported in Fig. 4(a). By employing the Nelder-Mead algorithm, an approximation of the optimal PV amount was found to be 262.4 kW. Compared to the costs when no PVs are installed, there is an approximate 14.5% decrease in the total costs of the WDN. The total cost exhibits a quadratic trend with noticeable fluctuations, particularly in the vicinity of the approximate optimal value. It is worth noting that the clear

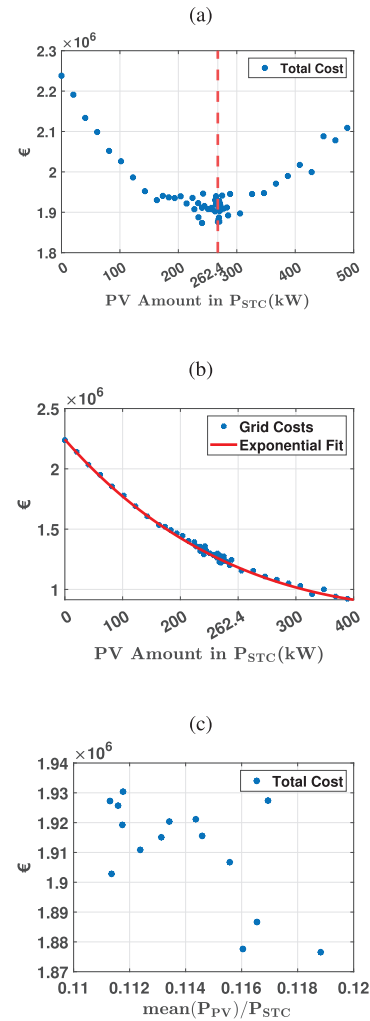


Fig. 4. (a) Total cost of the WDN over the lifespan $\ell_{\text{pv}} = 25$ of the PV panels including the installation and maintenance costs of the PV panels and the electricity costs of the WDN for a given amount of PVs, $J_c(x) + \ell_{\text{pv}}J_o(x)$. (b) Total costs of power bought from the grid and an exponential function fit to the data. (c) Total cost of the WDN and the ratio of mean PV production of sampled yearly PV profile and the installed PV amount.

trend in cost values supports our expectation that a one-year PV sample should give us a reasonably close estimate of the expected cost. The fluctuations in the cost values stem from the inherent noise introduced by the yearly PV profiles, which are generated through sampling from stochastic variables. These fluctuations around the approximate optimal value are investigated in Fig. 4(c). The figure shows how the total cost of the network is associated with the ratio of mean PV power production from the yearly sampled PV profile to the installed PV amount. Specifically, it focuses on PV amounts that are close to the approximate optimal value. Notably, higher mean PV profiles are associated with relatively lower costs, while lower mean PV profiles result in higher costs. The costs of purchasing power from the grid are given in Fig. 4(b) and approximated with an exponential function fit to the data. This is the only stochastic term in the total network cost $J_c(x) + \ell_{\text{pv}}J_o(x)$ since installation and maintenance costs are deterministic functions of the installed PV amount. Assuming the fit exponential function represents the expected electricity

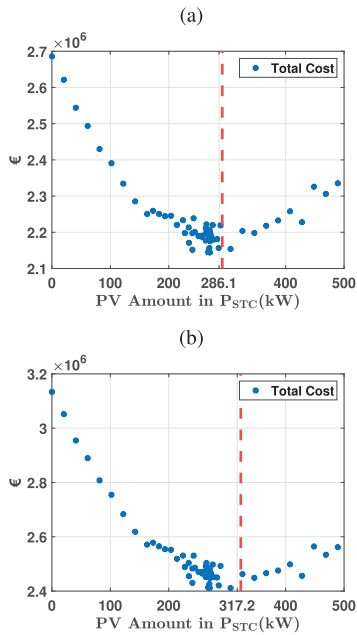


Fig. 5. Total costs for different lifespans of PVs (a) $\ell_{pv} = 30$, (b) $\ell_{pv} = 35$. The red dashed lines represent the approximated optimal quantity of PV panels, obtained by fitting an exponential function to network costs and minimizing the expected total cost based on this exponential function.

costs, meaning $J_o(x)$ is considered deterministic and given by the summation of PV installation costs and the exponential fitting of the electricity costs, the optimal PV amount is estimated to be 246.1 kW. Now, there are two values: one returned by the Nelder-Mead method and another from the exponential fitting of electricity costs. Nelder-Mead method quickly gets close to the vicinity of optimal value, but it does not converge to the optimal value for stochastic optimization problems. Therefore, the value obtained through the Nelder-Mead method can only be considered to be in the vicinity of the optimal value. On the other hand, the value obtained by fitting an exponential function to the grid costs, as shown in Fig. 4(b), may be deemed more reliable, given the evident exponential trend in the grid costs. Nonetheless, for enhanced precision in results, additional evaluations of total WDN costs under varying amounts of PVs should be conducted.

The exponential fit of the network operation costs enables an investigation into the impact of using a non-constant PV efficiency, λ_{pv} . Assuming this fit represents the expected electricity costs for a given PV panel quantity, the total expected network costs can be calculated with a yearly linear decline in PV efficiency, using the exponential fit to calculate operational network costs for each year's PV efficiency. The PV size can then be optimized using this yearly decline model. The optimal PV size determined through this approach is 245.4 kW, which is nearly identical to the result obtained with a constant PV efficiency, where the optimal PV size is found to be 246.1 kW. This indicates that the assumption of constant PV efficiency has little impact on the result in this simulation.

The costs for different life spans $\ell_{pv} = 30, \ell_{pv} = 35$ are given in Fig. 5. As the life span increases, we anticipate that the optimal PV amount will exceed the value observed in

the $\ell_{pv} = 25$ case. This expectation arises from the fact that with a longer life span, OPEX $J_o(x)$ increases at a slower rate for higher PV quantities, as they can generate more free power for the WDN. On the other hand, CAPEX $J_c(x)$ remains constant. This phenomenon can be observed in the figures, where the cost for larger PV amounts approaches the cost of the approximate optimal amount (262.4 kW) for $\ell_{pv} = 25$ as the life span increases. Consequently, the approximate optimal PV amount shifts toward higher values for longer life spans. If we again use the exponential function fit to the network costs and use it as the expected network cost, the approximate optimal PV amount for $\ell_{pv} = 30$ increases to 286.1 kW, and for $\ell_{pv} = 35$, it further rises to 317.2 kW. Under this assumption, the cost reduction achieved with $\ell_{pv} = 30$ is 18.1%, while for $\ell_{pv} = 35$, it reaches 21.4%.

The simulations are conducted using the approximate linear model (22b) rather than the EPANET model. To assess the impact of using the linear model, we conducted a simulation study comparing the operational costs obtained from the linear model and the EPANET model. The simulations are performed for 20 different PV capacities, ranging from 20 to 400 kW, using 20 representative PV production sample days to capture annual variability. An affine function is fit to express the operational costs of the EPANET model in terms of the linear model's costs. By combining this with the exponential fit for network costs from the linear model, the operational costs of the WDN using the EPANET model are approximated. The optimized PV capacity from this approximation is 242.7 kW, which is very close to the 246.1 kW obtained using the exponential fit alone, with only a 1.3% difference. These results indicate that the linear model provides a reliable approximation of the operational costs obtained from the EPANET model.

It is important to note that the calculation of the approximate optimal PV amount is based on certain assumptions. We used electricity prices from 2022 and assumed a constant water demand throughout the lifespan of the PVs. Additionally, we did not consider the potential impact of climate change on PV power production and water demand. However, these assumptions were applied consistently across all PV amounts, enabling us to identify the approximate optimal PV amount and achieve cost reduction with these assumptions. Moving forward, more accurate representations of water demand and electricity prices are needed to obtain more reliable results. Additionally, a constant PV efficiency model $P_{STC} = \lambda_{pv}x$ is used to ease the computations. Without this simplification, the results would have shown slightly less overall savings. That is because of the convex-like shape of the grid costs [see Fig. 4(b)]. To illustrate this, consider the value returned by the Nelder-Mead method: 262.4 kW. Now, assume that PV efficiency is not constant but $P_{STC} = 270$ kW for the first half of the PV lifespan and $P_{STC} = 250$ kW for the second half. Looking at Fig. 4(b), it can be seen that the total cost in this situation would be more than if we stuck with a constant $P_{STC} = 262.4$ kW throughout. Furthermore, a constant cost for installing PVs is used based on the market price for PV panels. However, in reality, this likely makes the overall cost estimate higher than it would be. Usually,

when more PV panels are added, the cost per panel goes down, but the model used in this article does not consider that, so it might have overestimated the total costs. A one-year PV sample is used to simulate the WDN and approximate the total cost (37) for computational purposes. Despite the potential for some inaccuracies, examining Fig. 4(b), a clear trend in grid costs can be seen. Also, the differences in costs for similar PV quantities are quite small. Therefore, it can be said that using one-year samples did not lead to significant inaccuracies. Another factor affecting the results is the half-full requirements on the tanks. With a lower bound on tank levels, pump schedules could have been more flexible, making the WDN operation more cost-effective. This would have resulted in a lower total cost for each given PV amount. However, the impact of this on the optimal PV quantity is not straightforward, as it is challenging to estimate the precise cost reduction for each PV amount.

VII. CONCLUSION

We have introduced a method to find the optimal amount of PVs to install for a WDN along with a probabilistic PV power production model. The results show that by installing the optimal number of PV systems in the Randers WDN, it is possible to achieve a reduction of 14.5% in network costs. Note that this number is obtained by assuming the pumps of the WDN will be scheduled with the presented stochastic controller throughout the lifespan of PVs. However, an alternative control method could easily be integrated into the procedure presented in this article by simulating the WDN with the desired control method. We might question whether assuming a consistent pump scheduling method throughout the lifespan of PVs is practical. This leads to concerns about the reliability of the determined PV amounts if, for example, the WDN facility plans to use different pump scheduling methods. However, as long as the facility adopts a pump scheduling method that minimizes the economic costs of pumps, the resulting pump schedules should not significantly differ from those of other control methods. This is because the primary drivers of cost minimization are the available PV power and grid electricity prices, both of which remain independent of the specific pump scheduling method. Consequently, the proposed PV amounts can be relied upon, provided the WDN facility uses similar pump scheduling methods aimed at minimizing electricity costs for the pumps. Although the presented procedure provides a detailed approach to determining the required PV installation, it still requires additional steps such as fitting the linear model of the network and simulating the WDN which is time-consuming. It would be desirable to have a simpler method for determining the number of PVs to be installed. However, this is not feasible by solely analyzing the simulation results of a single WDN, as the optimal PV amount depends on various factors, including elevated tank and water pump capacities in the network, average daytime, and even the seasonal variation of daytime throughout the year. To obtain a simplified formula for PV installation based on these parameters, the presented procedure could be applied to different WDNs with varying characteristics in terms of these variables. Then, a regression model could be fit to optimal

PV amount as a function of certain WDN properties, climate properties of WDN location, etc. This would enable estimating the optimal amount of PVs without having run the WDN simulations. However, to make this work, we need a fast way to calculate or estimate the total costs of the WDN as stated in (34). One way to do this is by using a neural network to mimic the input-output characteristic of the pump scheduling problem (22), as running a neural network is a lot faster than solving the given optimization problem. Our initial attempts at training a neural network yielded suboptimal approximations of the pump scheduling problem's input-output characteristics. Further exploration in this direction could bring other potential advantages as well. If successful, a well-trained neural network could accelerate simulations, offering potential benefits such as estimating the total cost of the network (34), without relying on the assumption of constant PV efficiency. This, in turn, could lead to more precise approximations of the overall cost function. Another direction for future work is evaluating the robustness of the proposed method under different PV power generation models by comparing network simulations using synthetic power data and data generated from the proposed PV power generator and from various other PV power models. This will help assess the influence of the PV power sample generator on network costs. Additionally, simulations with various optimization-based pump scheduling methods will be conducted to analyze their impact on the network's operational costs and performance. As another future work, the linear network model (21) could also be improved with the help of demand allocation methods that estimate the demands of each node. If the demand information of each node can be integrated into the linear model (21) without making the model complex, this could improve the accuracy of the model without increasing the computational complexity of the controller calculation.

ACKNOWLEDGMENT

The authors acknowledge Verdo Company, Peter Nordahn, and Steffen Schmidt for providing them with the EPANET model and the network information.

REFERENCES

- [1] M. N. Sharif, H. Haider, A. Farahat, K. Hewage, and R. Sadiq, "Water-energy Nexus for water distribution systems: A literature review," *Environ. Rev.*, vol. 27, no. 4, pp. 519–544, Dec. 2019.
- [2] Y. Guo, S. Wang, A. F. Taha, and T. H. Summers, "Optimal pump control for water distribution networks via data-based distributional robustness," *IEEE Trans. Control Syst. Technol.*, vol. 31, no. 1, pp. 114–129, Jan. 2023.
- [3] C. S. Kallesøe, T. N. Jensen, and J. D. Bendtsen, "Plug-and-Play model predictive control for water supply networks with storage," *IFAC-PapersOnLine*, vol. 50, no. 1, pp. 6582–6587, Jul. 2017.
- [4] Q. Zhao, W. Wu, A. R. Simpson, and A. Willis, "Water distribution system optimization considering behind-the-meter solar energy with a hydraulic power-based search-space reduction method," *J. Water Resour. Planning Manage.*, vol. 149, no. 9, Sep. 2023, Art. no. 04023046.
- [5] A. V. Mercedes Garcia, F. J. Sánchez-Romero, P. A. López-Jiménez, and M. Pérez-Sánchez, "A new optimization approach for the use of hybrid renewable systems in the search of the zero net energy consumption in water irrigation systems," *Renew. Energy*, vol. 195, pp. 853–871, Aug. 2022. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0960148122008928>

- [6] R. Cantu-Funes and L. C. Coelho, "Simulation-based optimization of pump scheduling for drinking water distribution systems," *Eng. Optim.*, vol. 55, no. 5, pp. 841–855, May 2023.
- [7] D. C. L. Ormsbee and S. Lingireddy. (2009). *Optimal Pump Scheduling for Water Distribution Systems*. [Online]. Available: <http://www.uky.edu/WDST/PDFs/>
- [8] M. Castro-Gama, Q. Pan, E. A. Lanfranchi, A. Jonoski, and D. P. Solomatine, "Pump scheduling for a large water distribution network. Milan, Italy," *Proc. Eng.*, vol. 186, pp. 436–443, Jan. 2017.
- [9] K. M. H. Baunsgaard, O. Ravn, C. S. Kallesøe, and N. K. Poulsen, "MPC control of water supply networks," in *Proc. Eur. Control Conf. (ECC)*, Jun. 2016, pp. 1770–1775.
- [10] K. S. Ayyagari, S. Wang, N. Gatsis, A. F. Taha, and M. Giacomoni, "Co-optimization of interdependent water and power distribution networks," in *Proc. IEEE Power Energy Soc. Innov. Smart Grid Technol. Conf. (ISGT)*, Feb. 2021, pp. 1–5.
- [11] Y. Wang, T. Alamo, V. Puig, and G. Cembrano, "Economic model predictive control with nonlinear constraint relaxation for the operational management of water distribution networks," *Energies*, vol. 11, no. 4, p. 991, Apr. 2018. [Online]. Available: <https://www.mdpi.com/1996-1073/11/4/991>
- [12] F. Fiedler, A. Cominola, and S. Lucia, "Economic nonlinear predictive control of water distribution networks based on surrogate modeling and automatic clustering," *IFAC-PapersOnLine*, vol. 53, no. 2, pp. 16636–16643, 2020.
- [13] A. P. Goryashko and A. S. Nemirovski, "Robust energy cost optimization of water distribution system with uncertain demand," *Autom. Remote Control*, vol. 75, no. 10, pp. 1754–1769, Oct. 2014.
- [14] X. Zheng, Y. Wang, E. Weyer, and C. Manzie, "Economic model predictive control of water distribution systems with solar energy and batteries," in *Proc. 62nd IEEE Conf. Decis. Control (CDC)*, Dec. 2023, pp. 265–270.
- [15] M. Ürkmez, C. Kallesøe, J. D. Bendtsen, and J. Leth, "Scenario based cost optimization of water distribution networks powered by grid-connected photovoltaic systems," in *Proc. IEEE Conf. Control Technol. Appl. (CCTA)*, Aug. 2023, pp. 810–816.
- [16] T. Huld et al., "A power-rating model for crystalline silicon PV modules," *Sol. Energy Mater. Sol. Cells*, vol. 95, no. 12, pp. 3359–3369, Dec. 2011. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0927024811004442>
- [17] F. E. Atencio Espejo, S. Grillo, and L. Luini, "Photovoltaic power production estimation based on numerical weather predictions," in *Proc. IEEE Milan PowerTech*, Jun. 2019, pp. 1–6.
- [18] S. Theodorides, G. Makrides, A. Livera, M. Theristis, P. Kaimakis, and G. E. Georgiou, "Day-ahead photovoltaic power production forecasting methodology based on machine learning and statistical post-processing," *Appl. Energy*, vol. 268, Jun. 2020, Art. no. 115023. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0306261920305353>
- [19] A. Mellit, A. M. Pavan, and V. Lughi, "Short-term forecasting of power production in a large-scale photovoltaic plant," *Sol. Energy*, vol. 105, pp. 401–413, Jul. 2014.
- [20] J. Á. González Ordiano, S. Wacziarg, M. Reischl, R. Mikut, and V. Hagenmeyer, "Photovoltaic power forecasting using simple data-driven models without weather data," *Comput. Sci.-Res. Develop.*, vol. 32, nos. 1–2, pp. 237–246, Mar. 2017.
- [21] P. Bacher, H. Madsen, and H. A. Nielsen, "Online short-term solar power forecasting," *Solar Energy*, vol. 83, no. 10, pp. 1772–1783, 2009.
- [22] Y. Xia, J. Wang, Z. Zhang, D. Wei, and L. Yin, "Short-term PV power forecasting based on time series expansion and high-order fuzzy cognitive maps," *Appl. Soft Comput.*, vol. 135, Mar. 2023, Art. no. 110037. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1568494623000558>
- [23] A. Starosta, K. Kaushik, P. Jhaveri, N. Munzke, and M. Hiller, "A comparative analysis of forecasting methods for photovoltaic power and energy generation with and without exogenous inputs," in *Proc. 38th Eur. Photovoltaic Sol. Energy Conf. Exhib.*, 2021, p. 938.
- [24] I. Khenissi, M. A. Fakhfakh, R. Sellami, and R. Neji, "A new approach for optimal sizing of a grid connected PV system using PSO and GA algorithms: Case of Tunisia," *Appl. Artif. Intell.*, vol. 35, no. 15, pp. 1930–1951, Dec. 2021.
- [25] Y. Zhang, T. Ma, P. Elia Campana, Y. Yamaguchi, and Y. Dai, "A techno-economic sizing method for grid-connected household photovoltaic battery systems," *Appl. Energy*, vol. 269, Jul. 2020, Art. no. 115106. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0306261920306188>
- [26] V. Janamala and K. R. Rani, "Optimal allocation of solar photovoltaic distributed generation in electrical distribution networks using archimedes optimization algorithm," *Clean Energy*, vol. 6, no. 2, pp. 271–287, Apr. 2022.
- [27] H. L. Thadani and Y. I. Go, "Large-scale solar system design, optimal sizing and techno-economic-environmental assessment," *Sustain. Energy Res.*, vol. 10, no. 1, pp. 1–22, Aug. 2023.
- [28] M. Duong, T. Pham, T. Nguyen, A. Doan, and H. Tran, "Determination of optimal location and sizing of solar photovoltaic distribution generation units in radial distribution systems," *Energies*, vol. 12, no. 1, p. 174, Jan. 2019.
- [29] M. A. Pardo and F. J. Navarro-González, "Sizing and scheduling optimisation method for off-grid battery photovoltaic irrigation networks," *Renew. Energy*, vol. 221, Feb. 2024, Art. no. 119822. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0960148123017378>
- [30] T. Castillo-Calzadilla, C. M. Andonegui, M. Gómez-Goiri, A. M. Macarulla, and C. E. Borges, "Systematic analysis and design of water networks with solar photovoltaic energy," *IEEE Trans. Eng. Manag.*, vol. 69, no. 3, pp. 628–638, Jun. 2022.
- [31] R. Zieba Falama, H. Bakari, and V. Dumbrava, "Double-objective optimization-based firefly algorithm of a stand-alone photovoltaic/water pumping system for water supply in rural and remote areas: A case study," *J. Electr. Syst. Inf. Technol.*, vol. 8, no. 1, pp. 1–25, Dec. 2021.
- [32] B. Bouzidi, "New sizing method of PV water pumping systems," *Sustain. Energy Technol. Assessments*, vol. 4, pp. 1–10, Dec. 2013. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S221313881300057X>
- [33] M. Ürkmez, C. S. Kallesøe, J. D. Bendtsen, and J. Leth, "Day-ahead PV power forecasting for control applications," in *Proc. 48th Annu. Conf. IEEE Ind. Electron. Soc.*, pp. 1–6, Oct. 2022.
- [34] D. Faïman, "Assessing the outdoor operating temperature of photovoltaic modules," *Prog. Photovolt., Res. Appl.*, vol. 16, no. 4, pp. 307–315, Jun. 2008.
- [35] C. F. Lo, "WKB approximation for the sum of two correlated lognormal random variables," *Appl. Math. Sci.*, vol. 7, no. 128, pp. 6355–6367, Aug. 2013.
- [36] C. Kallesøe, J. Aarestrup, and K. Rokkjær, "Energy optimization for booster sets," *World Pumps*, vol. 2011, no. 12, pp. 24–30, Dec. 2011. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0262176211703662>
- [37] T. J. Broomhead, C. Manzie, R. C. Shekhar, and P. Hield, "Robust periodic economic MPC for linear systems," *Automatica*, vol. 60, pp. 30–37, Oct. 2015. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0005109815002691>
- [38] S. Jerez et al., "The impact of climate change on photovoltaic power generation in Europe," *Nature Commun.*, vol. 6, no. 1, p. 10014, Dec. 2015.
- [39] J. A. Nelder and R. Mead, "A simplex method for function minimization," *Comput. J.*, vol. 7, no. 4, pp. 308–313, 1965.
- [40] J. C. Lagarias, J. A. Reeds, M. H. Wright, and P. E. Wright, "Convergence properties of the nelder-mead simplex method in low dimensions," *SIAM J. Optim.*, vol. 9, no. 1, pp. 112–147, Jan. 1998.
- [41] E. Barykina and A. Hammer, "Modeling of photovoltaic module temperature using faïman model: Sensitivity analysis for different climates," *Sol. Energy*, vol. 146, pp. 401–416, Apr. 2017.
- [42] E. Annigoni, A. Virtuani, M. Cacciavo, G. Friesen, D. Chianese, and C. Ballif, "35 years of photovoltaics: Analysis of the TISO-10-kW solar plant, lessons learnt in safety and performance—Part 2," *Prog. Photovolt., Res. Appl.*, vol. 27, no. 9, pp. 760–778, Sep. 2019.
- [43] A. Sangwongwanich, Y. Yang, D. Sera, and F. Blaabjerg, "Lifetime evaluation of grid-connected PV inverters considering panel degradation rates and installation sites," *IEEE Trans. Power Electron.*, vol. 33, no. 2, pp. 1225–1236, Feb. 2018.
- [44] R. Wiser, M. Bolinger, and J. Seel, "Benchmarking utility-scale PV operational expenses and project lifetimes: Results from a survey of U.S. solar industry professionals," Lawrence Berkeley Nat. Lab., Berkeley, CA, USA, Tech. Rep., 2020.