Forecasting Weather and Energy Demand for Optimization of Renewable Energy and Energy Storage Systems for Water Desalination

Om Sanan
Scarsdale High School
Scarsdale, USA
om.sanan007@gmail.com

Joshua Sperling
NREL
Golden, USA
joshua.sperling@nrel.gov

David Greene
NREL
Golden, USA
david.greene@nrel.gov

Ross Greer
UC San Diego
La Jolla, USA
regreer@ucsd.edu

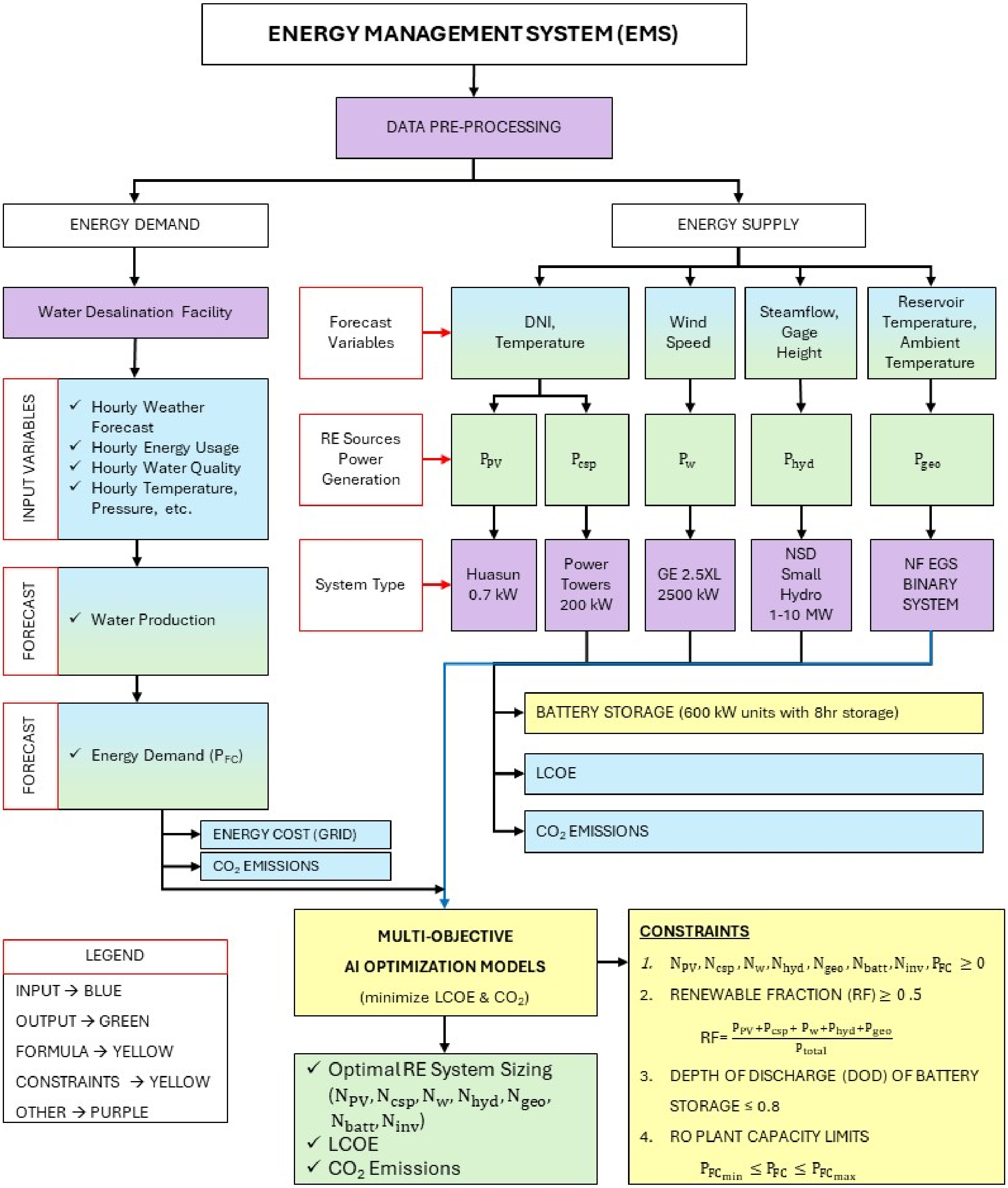
***Abstract*—The increasing intensity and frequency of water scarcity, carbon emissions, and climate risks pose critical challenges necessitating increased uptake of and a paradigm shift to energy- and climate-smart water desalination processes. This study employs metrics and a decision framework to enable and accelerate the energy efficiency, decarbonization, and cost-effectiveness of water desalination processes. As an essential step, we analyze various Renewable Energy (RE) sources, such as photovoltaic, wind, concentrated solar power, geothermal, and hydro energy; in addition, we examine battery storage systems to address the intermittency challenges associated with solar and wind energy. The feasibility of these diverse RE systems was assessed at four (4) mid-to-large scale U.S. desalination plants using operating plant and weather/environmental data, establishing optimization functions and constraints. In this research, to facilitate a comprehensive Energy Management System (EMS), we align RE generation with the anticipated energy demand of the plants. Machine Learning (ML) models, including SARIMA, Random Forest, XGBoost, and Gradient Boosting, are employed for forecasting water production, energy consumption, and long-term weather. The results show that Artificial Intelligence (AI) models, notably Gradient Boosting and an innovative XGBoost average method, demonstrated high accuracy in forecasting critical variables for RE systems in water desalination, with a normalized Root Mean Square Error of less than 10% for key metrics. This study can serve as a basis to optimize the mix of hybrid RE systems to minimize cost and carbon emissions.**

***Index Terms*—water, desalination, reverse osmosis, time series forecasting, renewable energy, optimization, energy management, decarbonization, weather forecasting, artificial intelligence.**

I. Introduction

The water-energy nexus represents a critical intersection in our sustainable future, highlighting the intricate link between water and energy efficiency, where each resource's management directly influences the availability and sustainability of the other [1]. In the face of a mounting global water crisis exacerbated by population growth, climate change, and unsustainable water management practices, water desalination has emerged as a critical technology to meet the escalating demand for freshwater resources [2]. Reverse Osmosis (RO) has been the most effective desalination process with an approximate 65% market share [3], but its costs (especially energy costs) need to reduce further to make it commercially viable and environmentally sustainable. RO's Levelized Cost of Water (LCOW) and Specific Energy Consumption (SEC) have been declining due to engineering innovation, high-efficiency pressure pumps, improvement in RO membrane structure and use of energy recovery devices [4], [5]. As an example, the SEC has reduced from 8 kWh/m^3 in the 1970s to 2.5-3.0 kWh/m^3 approaching the thermodynamic limit of approximately 1 kWh/m^3 [2]. Renewable energy offers the potential to further lower the costs of water desalination, and substantially reduce emissions. Even though the International Renewable Energy Agency (IRENA) claims that RE could power almost all the world's energy needs by 2050, highlighting the vast potential of these sources for sustaining water desalination processes, today less than 1% of the water desalination plants' energy need globally is met by RE [6]. Recently there has been significant interest (and studies) in implementing RE systems to power RO desalination; however, these systems have generally been small/pilot-scale. There are a few notable exceptions that prove that with the right regulations and structure, 100% RE transition can be achieved. For example, the Al Khafji Seawater Reverse Osmosis (SWRO) plant in Saudi Arabia, operational in 2018, produces 16 MGD water and is the first-ever large-scale desalination plant connected to a grid-tied polycrystalline silicon PV-RO system that is 100% powered by RE [7]. Its solar PV arrays have an output of 20 MW, are located 1 km away from the SWRO plant and cover an area of 900,000m^2. Similarly, the Perth SWRO plant in Australia with a capacity of 36 MGD sources electricity from 48 wind turbines at Emu Downs Wind Farm (electricity is first sold to the utility) with 83 MW of capacity and located 260 km north of the city [5] validating that land considerations for onsite RE plants can be overcome.

We present a novel solution for decarbonizing water desalination by optimizing RE systems using AI by analyzing real-world data from four U.S. water desalination plants at Tampa Bay in Florida, San Antonio in Texas, Alameda County in California, and Kay-Bailey Hutchison in Texas. Our research employs metrics and a decision framework to decarbonize desalination process by 50%, 75% and 100% over time by using a mix of renewable energy sources such as Photovoltaic


ENERGY MANAGEMENT SYSTEM (EMS)
DATA PRE-PROCESSING
ENERGY DEMAND
ENERGY SUPPLY
Water Desalination Facility
INPUT VARIABLES
✓ Hourly Weather Forecast
✓ Hourly Energy Usage
✓ Hourly Water Quality
✓ Hourly Temperature, Pressure, etc.
FORECAST
✓ Water Production
FORECAST
✓ Energy Demand (P_{FC})
ENERGY COST (GRID)
CO_2 EMISSIONS
Forecast Variables
DNI, Temperature
Wind Speed
Steamflow, Gage Height
Reservoir Temperature, Ambient Temperature
RE Sources Power Generation
P_{PV}
P_{csp}
P_w
P_{hyd}
P_{geo}
System Type
Huasun 0.7 kW
Power Towers 200 kW
GE 2.5XL 2500 kW
NSD Small Hydro 1-10 MW
NF EGS BINARY SYSTEM
BATTERY STORAGE (600 kW units with 8hr storage)
LCOE
CO_2 EMISSIONS
MULTI-OBJECTIVE AI OPTIMIZATION MODELS
(minimize LCOE & CO_2)
✓ Optimal RE System Sizing ($N_{PV}, N_{csp}, N_w, N_{hyd}, N_{geo}, N_{batt}, N_{inv}$)
✓ LCOE
✓ CO_2 Emissions
CONSTRAINTS
1. $N_{PV}, N_{csp}, N_w, N_{hyd}, N_{geo}, N_{batt}, N_{inv}, P_{FC} \geq 0$
2. RENEWABLE FRACTION (RF) ≥ 0 .5
$RF = \frac{P_{PV} + P_{csp} + P_w + P_{hyd} + P_{geo}}{P_{total}}$
3. DEPTH OF DISCHARGE (DOD) OF BATTERY STORAGE ≤ 0.8
4. RO PLANT CAPACITY LIMITS
$P_{FC_{min}} \leq P_{FC} \leq P_{FC_{max}}$
LEGEND
INPUT → BLUE
OUTPUT → GREEN
FORMULA → YELLOW
CONSTRAINTS → YELLOW
OTHER → PURPLE

Footnotes:

1) P_{FC} is the energy demand of desalination plant.

2) P_{pv}, P_{csp}, P_w, P_{hyd}, P_{geo} represent renewable energy generation from PV, CSP, WT, HT and GE, respectively.

3) N_{pv}, N_{csp}, N_w, N_{hyd}, N_{geo}, N_{batt}, N_{inv} represent the number of PV arrays, CSP units, wind turbines, hydro turbines, geothermal power plants, battery banks and inverters, respectively.

Fig. 1: Energy Management System

Energy (PV), Wind Energy (WE), Concentrated Solar Power (CSP), Geothermal Energy (GE) and Hydro Power (HP). While RE will play an integral part in the decarbonization of the desalination and the electric grid, no single RE source can service the entire need as the RE power source is intermittent [8]. A combination of hybrid RE will be required to maintain reliability at the utility level, including baseload (minimum power load requirement) with geothermal, hydro, biomass, nuclear), intermediate load (CSP with storage, hydro), and peak load (PV, wind, CSP without storage) [9].

Our proposed EMS framework shown in Fig. 1 can be used to generate optimal sizing of each RE system to satisfy the desalination plants' energy demand. The EMS first learns by assessing 5 years of historical daily/hourly operational data from the desalination plant including water production, energy consumption, water quality and 10-year historical hourly weather data to establish optimization functions and constraints. Various ML models, including Seasonal Autoregressive Integrated Moving Average (SARIMA) [10], Random Forest, XGBoost, and Gradient Boosting, are employed for weather and energy load forecasting. We then estimate the energy production potential of each RE source, factoring in surplus or deficiency management through battery storage and utility grid interconnection.

Decarbonization of water desalination will require a diversified approach, incorporating water reuse/recycle, water and electricity demand reduction, varying mix of renewables, oversized RE capacities to overcome high cost of long-duration storage, energy efficiency measures using digital technologies, and demand side management.

II. Related Research

The existing literature lacks a comprehensive approach to optimizing the diversity of grid-connected RE sources across different regions using real-life RO desalination plants to minimize the energy cost and carbon footprint.

Several studies have been conducted to forecast water production, energy demand of desalination, and weather using ML models. However, most of these focus on short-term forecasting. [3] designed a hybrid RE system coupled with RO desalination using solar and wind energy, battery back-up and multi-criteria AI optimization models. [11] used an off-grid power system to power a SWRO pilot plant addressing optimal sizing of off-grid SWRO plants. [12] demonstrated the potential of ML to forecast solar energy generation and improve utilization in desalination systems. [13] conducted a survey for use of ML models for applications in weather and climate forecasting to bridge the gap between short-term and long-term weather forecasting. [14] discussed the development and application of ARIMA model for weather forecasting, specifically focusing on visibility forecasting. [15] presented a Vector Autoregression (VAR) weather model to forecast key weather variables (temperature, solar radiation, and wind speed) for electricity supply and demand in 61 U.S. cities. These studies are a subset of our scope to decarbonize desalination. Additionally, most data used in these studies was either simulated or on a small scale.

III. Analysis of Renewable Energy Sources

Decarbonizing water desalination necessitates examining the feasibility, efficiency, and integration potential of different RE sources, including photovoltaic panels, wind turbines, concentrated solar power, geothermal units, and hydro turbines. Through predictive analytics and data-driven insights, we aim to craft a nuanced understanding of how each energy source can be harnessed effectively, considering the environmental conditions and technological capabilities at our disposal. Our focus extends beyond mere adoption of RE sources; it encompasses a thorough evaluation of factors that influence the efficiency, deployment, and performance of these systems.

A. Power Estimation Formulas and Assumptions

In order to find an optimal combination of RE sources for each site, we must first outline all the factors that play a role in this in order to analyze the effectiveness of each RE source individually. More specifically, we have outlined a framework consisting of a set of formulas, variables, and constraints.

The energy generation formulas for each RE source depend on weather/environmental factors, including temperature (for PV and CSP), solar irradiance (for PV and CSP), wind speed (for wind), streamflow (aka discharge) and gage height (for hydro) as shown in Table I [4], [16]. We have used state-of-the-art technologies for RE systems currently available (and expect further ongoing improvements), which will enhance energy generation, and lower the Levelized Cost of Energy (LCOE). Table II shows the assumptions underlying the energy estimation and battery storage formulas. In our analysis, we used Huasun Solar - Himalaya G12 Series HJT PV solar module; GE 2.5XL wind turbine with a 2,500 kW rated capacity; power tower with 10-hour storage for CSP; run-of-river hydro (1 - 10 MW) and mini hydro (100 kW-1 MW) power plant without storage; Next Frontier Enhanced Geothermal Service (EGS) using Binary cycle technology with 1.5 km well; and 600 kW power rated battery storage with 8-hour discharge time.

We examine battery storage to address intermittency challenges associated with RE systems: 8-hour battery storage units are used for system sizing, with excess RE production (once fully charged) sold to the utility. The battery discharges if the energy consumption is greater than RE generation. Scenarios both with and without battery storage are evaluated to assess the cost.

B. Levelized Cost of Energy and Carbon Emissions

The LCOE, which is the per-unit cost of generating electricity over the entire life cycle, for each RE source was obtained from the National Renewable Energy Laboratory (NREL) Annual Technology Baseline (ATB) workbook [17]. Table III shows the LCOE and carbon emissions for each of the four desalination plants reviewed. The LCOE of RE has reduced significantly over the last decade – per IRENA from

2010 to 2022, the global LCOE of PV has reduced by 89% to $0.049/kWh, onshore wind by 69% to $0.033/kWh, and CSP by 69% to $0.118/kWh [18] – and in many cases lower than the cost from the grid.

TABLE I: Power Estimation Formulas

Formula/Assumption	Description
$P_{\text{pv}} = \eta_{\text{inv}} \cdot \eta_{\text{B}} \cdot \eta_{\text{r}} \cdot T_{\text{c}} \cdot A_{\text{PV}} \cdot I$	Photovoltaic Power Calculation
$T_{\text{c}} = (1 - \beta \cdot (T_{\text{cell}} - 25))$	Temperature Correction Factor
$T_{\text{cell}} = T_{\text{a}} + I \cdot \frac{(\text{NOCT}-20)}{800}$	Photovoltaic Cell Temperature Calculation
$P_{\text{w}} = \frac{0.5 \cdot \rho \cdot A \cdot v^3 \cdot C_{\text{p}}}{1000}$	Wind Power Calculation
$P_{\text{cs}} = A \cdot I \cdot \eta_{\text{sc}} \cdot \text{CF}$	Concentrated Solar Power Calculation
$P_{\text{h}} = \eta_{\text{h}} \cdot \rho \cdot g \cdot h \cdot Q/1000$	Hydro Power Calculation
$P_{\text{GT}} = \eta mc\Delta T$	Geothermal Power Calculation
$E_{\text{b}}(t) = E_{\text{b}}(t-1) + (P_{\text{pv}}(t) + P_{\text{w}}(t) + P_{\text{cs}}(t) + P_{\text{h}}(t) - \frac{PF(t)}{\eta_{\text{inv}}}) \cdot \eta_{\text{bch}}$	Battery Charging State (excess energy sold to the utility)
$E_{\text{b}}(t) = E_{\text{b}}(t-1) - (P_{\text{pv}}(t) + P_{\text{w}}(t) + P_{\text{cs}}(t) + P_{\text{h}}(t) - \frac{PF(t)}{\eta_{\text{inv}}}) \cdot \eta_{\text{bdch}}$	Battery Discharging State (demand > power generation)
$E_{\text{bmax}} = N_{\text{batt}} \cdot E_{\text{bsc}}$	Maximum Battery Energy Capacity
$E_{\text{bmin}} = (1 - \text{DOD}) \cdot E_{\text{bmax}}$	Minimum Battery Energy Capacity
$E_{\text{bmin}} \leq E_{\text{b}}(t) \leq E_{\text{bmax}}$	Battery Energy Capacity Constraint
$LCOE = \frac{(CRF*CAPEX+FOM)*1000}{(CF*8760)+VOM}$	Levelized Cost of Energy (LCOE)

TABLE II: Renewable Energy and Battery Storage Modeling Assumptions. Constants are in blue, variables are in black.

Parameter	Unit/Value	Parameter	Unit/Value
PV Panel Power Rating	700 W	**Hydro Power Rating**	Varies kW
η_{inv} (Inverter Efficiency)	95%	η_h (Efficiency of Hydro Turbine)	85%
η_B (Battery Efficiency)	100%	ρ (Water Density)	1,000 kg/m^3
η_r (Rated Solar Cell Efficiency)	22.50%	g (Acceleration Due to Gravity)	9.81 m/s^2
β (Temp. Coefficient of Efficiency)	-0.37%	h (Gage Height)	Varies m
Apv (Area of Each PV Module)	3.1 m^2	Q (Streamflow or Discharge of Water)	Varies m^3/s
I (Average Daily Solar Irradiance)	Varies kWh/m^2/day	CF (Capacity Factor Hydro)	62%
Ta (Ambient Temperature)	Varies °C		
NOCT (Nominal Operating Cell Temp)	44 °C	**Geothermal Power Rating**	Varies kW
		m (Mass Flow Rate of Geothermal Fluid)—	40 kg/s
Wind Turbine Power Rating	2,500 kW	c (Specific Heat Capacity of Fluid / Water)	4.186 kJ/(kg°C)
ρ(Air Density)	1.225 kg/m^3	$\Delta T = Tgr - Ta$ (Reservoir less Ambient Temp.)	Varies °C
A (Rotor Swept Area)	11,310 m^2	η_{gt} (Efficiency of Geothermal System)	90%
v (Wind Speed)	Varies m/s		
C_p (Power Coefficient)	35%	**Battery Storage**	
		Battery Power Capacity (BPC)	600 kW
CSP Power Output Capacity	200 kW	Discharge Time	8 hours
A (Area of Solar Collector)	4,047 m^2	Battery Storage Capacity (BSC)	4800 kWh
I (Average Daily Solar Irradiance)	Varies kWh/m^2/day	Depth of Discharge (DOD)	80%
η_{sc} (Efficiency of Solar Collector)	30.00%	η_{inv} (Inverter Efficiency)	95%
CF (Capacity Factor CSP)	45%	η_{bch} (Battery Charge Efficiency)	80%
		η_{bcch} (Battery Discharge Efficiency)	100%

TABLE III: LCOE and CO2 Emissions

LCOE ($/kWh)	**Utility**	**PV**	**Wind**	**CSP**	**Hydro**	**Battery**	**Grid CO2 Emissions (g/kWh)**
Tampa	$0.0730	$0.070	$0.038	$0.098	$0.080	$0.044	430
SAWS	$0.1010	$0.070	$0.054	$0.098	$0.080	$0.044	450
Alameda	$0.2120	$0.068	$0.054	$0.087	$0.080	$0.044	450
KBH	$0.0820	$0.068	$0.054	$0.077	$0.080	$0.044	330
Lifetime CO2 Emissions (g/kWh)		**43**	**13**	**28**	**21**	**33**	

IV. Forecasting Water Production, Energy Consumption, and Long-Term Weather

A. Forecasting Objective

This paper delves into a detailed analysis of the energy requirements of water desalination processes and the potential of various RE sources to meet these needs efficiently and sustainably. To properly model the RE energy generation for the future, we forecast the environmental conditions using AI modeling. Additionally, to determine the amount of RE system units needed to meet the energy demand for each plant, we used desalination plant energy consumption data.

B. Input Data and Output Quantities

As part of this study, we obtained operational data from 4 geographically diverse seawater and brackish water desalination facilities in the U.S. as shown in Table IV. With approximately 28 million gallons of actual water production per day (MGD), these plants offered a broad representation of the different environmental and operational conditions that could influence the feasibility and efficiency of RE systems for water treatment. The hourly/daily/monthly 5-year data from the facilities included water flows, hours in operation, energy use data, backwash water use, peak demand and quality metrics such as total dissolved solids (TDS), turbidity, temperature and pH. Fig. 2 shows the correlation between historical energy usage and treated water flow at San Antonio Water desalination plant. The plant data was pre-processed to account for variables like weather changes, plant shutdowns, and maintenance using statistical and AI tools, which formed the core of the approach. In addition, missing values and outliers were checked for and removed, other than the hydro data forecast (where was forward and backward filled).

TABLE IV: Summary Data from 4 Desalination Plants: Tampa Bay Desalination (TBD), San Antonio Water System (SAWS), Alameda County Water (ACW), and Kay Bailey Hutchison (KBH). Tampa Bay desalinates seawater and the others desalinate brackish water.

	TDS	Water Production		Energy Usage	SEC (avg.)
	ppm	MGD	m^3 / year	kWh / year	kWh / m^3
TBD, FL	35,000	8.2	11,288,134	43,023,680	3.81
SAWS, TX	1,325	3.9	5,349,394	5,019,000	0.94
ACW, CA	1,111	6.7	9,198,486	4,205,916	0.46
KBH, TX	2,500	9.0	12,433,762	22,380,772	1.8
Total	-	27.7	38,269,777	74,629,368	1.95

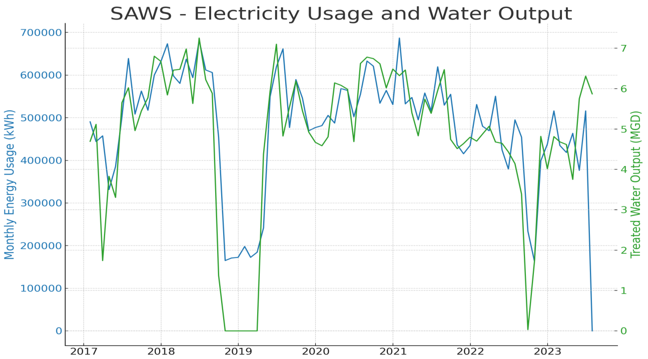


Fig. 2: SAWS - Historical Energy Usage vs. Treated Water

In addition, extensive weather data (irradiance, wind speed, temperature) as shown in Fig. 3 and environmental data (streamflow rate and gage height) for all study locations was collected, cleaned and processed to understand differences, anomalies, and ensure consistency. The The National Solar Radiation Database (NSRDB), created by NREL, was the primary database used, providing 10+ years hourly weather data from 2010 to 2021 [19]. Next, to evaluate run-of-river or mini hydro potential, we obtained streamflow and gage height data for a 4-year period (2018-2021) at 15-minute interval from USGS WaterWatch. The streams closest to the desalination plants were identified through the National Hydropower Asset Assessment Project (NHAAP) Public Portal [20].

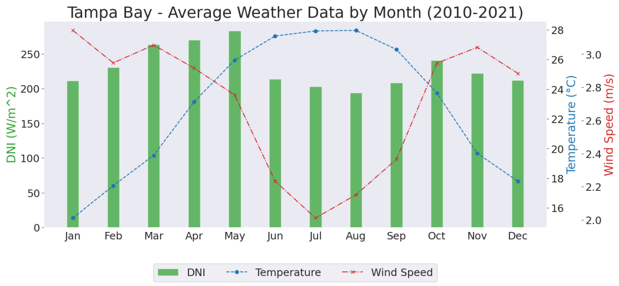


Fig. 3: Tampa Bay - Historical Monthly Weather

C. Methods

In this research, to obtain optimal sizing of each RE system to satisfy the desalination plants' energy demand, methods including predictive modeling for water production, energy consumption, and long-term weather/environmental forecasting have been used. The forecast modeling is split into two main areas. We first forecast the plant conditions, where we model each plant's treated water flows (Fig. 4) and energy consumption (Fig. 5) (historical data sourced from the respective water desalination plants), to use as constraints for RE generation potential on a given day. Second, we forecast environmental variables, which are further split into two sub-parts, due to the nature of the data obtained and the RE sources that were relevant. In the first sub-part, we forecast weather, specifically temperature (Fig. 6), DNI and wind speed (Fig. 7) (historical data sourced from NSRDB), which all serve as independent variables in the energy generation formulas of PV, CSP, and WE. Various ML models, including SARIMA, Random Forest, XGBoost, and Gradient Boosting, are employed for forecasting. In the second sub-part, the variables for hydro power generation, namely streamflow (the volume of water that pass through a hydroelectric power plant per unit time) and gage height (the height of water in a stream above a reference point) are forecasted. Fig. 8 shows streamflow (historical data sourced from USGS WaterWatch). These are also forecast using XGBoost.

1) Forecast of Energy Consumption and Water Flows: We employed a 60:20:20% train-test-validation dataset to ensure a robust evaluation of our model's performance. This split allowed us to have a sufficient number of samples for training, testing and validation while maintaining a balance between the different sets. Various models were then trained, including SARIMA, Random Forest, and Gradient Boosting Regressors. These models were chosen for their robustness in handling time series data and their ability to capture complex patterns in both water flow and energy consumption. An in-depth description of each model is provided.

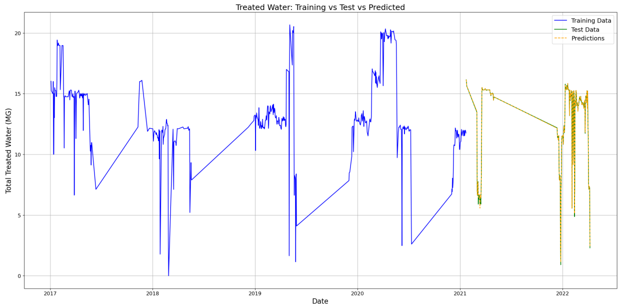


Fig. 4: Tampa Bay - Historical vs. Forecast Water Flow (MGD)

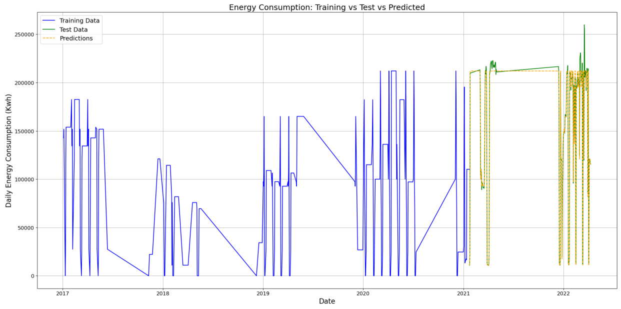


Fig. 5: Tampa Bay - Historical vs. Forecast Energy Consumption (kWh)

a) SARIMA: SARIMA is a time series forecasting model that accounts for seasonality in data [10]. SARIMA combines autoregressive (AR) and moving average (MA) components with differencing operations to make the time series stationary. The model introduces additional parameters to capture seasonal variations, including seasonal autoregressive (SAR) and

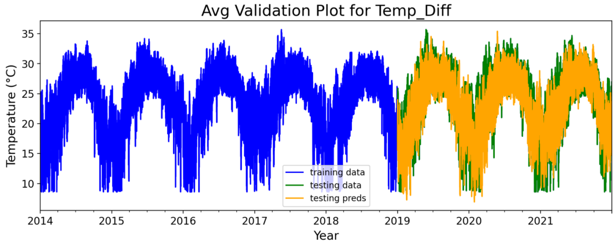


Fig. 6: Tampa Bay - Temperature (ºC) Train & Test Validation

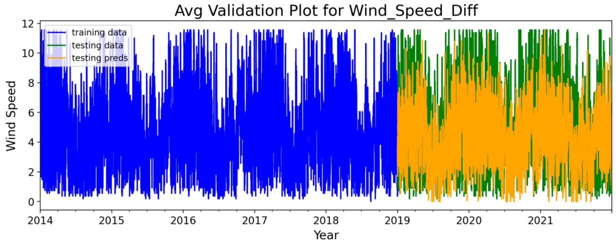


Fig. 7: Tampa Bay - Wind Speed (m/s) Train & Test Validation

seasonal moving average (SMA) terms. However, SARIMA assumes that the seasonality in the data follows a regular pattern with a constant length, which may not match the variably seasonal behavior of weather patterns.

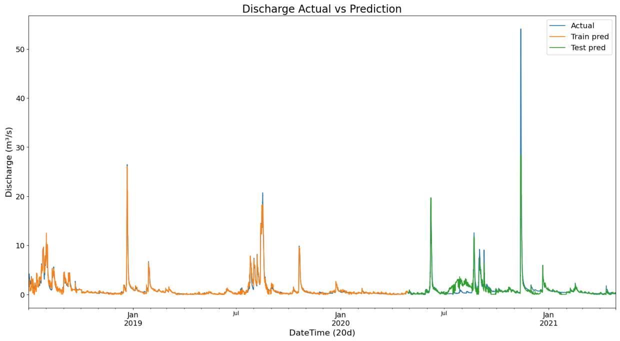


Fig. 8: Bullfrog Creek Tampa - Streamflow (m3/s) Train & Test Validation

b) Random Forest: Random Forest operates by constructing a multitude of decision trees during training and provides forecasts based on the average of the individual trees. What sets Random Forest apart is its incorporation of both bagging, a technique to reduce overfitting by training each tree on a random subset of the data, and random feature selection, enhancing model robustness and generalization. This ensemble approach allows Random Forest to excel in capturing complex relationships within data and handling high-dimensional feature spaces.

c) Gradient Boosting: Gradient Boosting is an ensemble learning technique that builds a strong predictive model by combining the forecasts of multiple weak models, typically decision trees. The core idea behind Gradient Boosting is to iteratively train new models to correct the errors made by the existing ensemble.

The training process involved fine-tuning the models with hyper-parameter optimization, using Grid Search Cross-Validation. This step was critical to identify the most effective parameters for each model for optimal performance. Parameters like Number of Estimators, Learning Rate, Max Depth, Min Samples Split, and Min Samples Leaf were adjusted to minimize validation error and enhance accuracy. Post-training, the models were evaluated using metrics such as RMSE (Root Mean Square Error), MAE (Mean Absolute Error), and Error Rate. The error rate function calculates the proportion of predictions that deviate from their actual values by more than a specified threshold. It is a measure of the frequency of significant errors in predictions, with a higher value indicating less accurate predictions. The best-performing models were then selected for each site and metric—Random Forest for all three metrics at Alameda, Gradient Boosting for Kay Bailey and San Antonio, and a combination of both for Tampa. The final stage involved making forecasts for the next 3 years, showcasing the models' capability to forecast trends and assist in strategic planning and decision-making.

2) Weather/Environmental Forecasting: The script created for this forecasting process is designed to interface with Streamlit, a popular framework for creating interactive web applications, primarily used here to fetch weather data from the NSRDB API and forecast future temperature, Direct Normal Irradiance (DNI), and wind speed locally. Since the wind speed obtained from NSRDB was recorded at a height of 2m, and our chosen wind turbines have a 139m hub height, we used the wind power law formula of [21] to recalibrate wind speed. The Streamlit interface allows users to input specific parameters like year and geographical coordinates, making the data fetching process interactive and user-friendly. The script then retrieves relevant weather data from the NSRDB API.

For forecasting, the script employs three different methods: XGBoost, Prophet, and a unique XGBoost average variant, as described below. The XGBoost model was initialized with an Autoregressive Forecaster from the skforecast library to easily facilitate time series forecasting.

a) XGBoost: XGBoost sequentially builds a multitude of weak learners, typically decision trees, with each subsequent tree aiming to correct the errors of its predecessors. It introduces an innovative approach by incorporating regularization terms in the objective function, which helps to prevent overfitting and enhances model generalization.

b) Prophet: Prophet is a decomposable time series forecasting model that is robust to missing data and shifts in trends [22]. It uses an additive approach that considers underlying components such as trends, seasonality, and holidays. The model adapts to the data's characteristics through a piecewise linear or logistic growth curve for trend forecasting, coupled with Fourier series for seasonal patterns, allowing flexibility in capturing seasonality changes over time. Prophet applies a

Bayesian framework to estimate uncertainties in the forecast, providing both point estimates and confidence intervals. This makes Prophet particularly suited for business forecasts with strong seasonal patterns and multiple seasonal cycles.

c) XGBoost Average Variant: The XGBoost average variant is a novel approach where the model first forecasts the differences between actual and average values of the parameters (temperature, DNI, wind speed) and then combines these forecasts with historical average data to make final forecasts. This method leverages the strengths of XGBoost in handling non-linear relationships and temporal dependencies, while also incorporating the historical average as a baseline, which often enhances the model's accuracy and robustness, especially in scenarios with cyclical or seasonal patterns.

The forecasting process starts with validating the models using a portion of the data to gauge their performance and accuracy. This structured approach ensures that the models are reliable and their forecasts are grounded in actual data trends. Additionally, the script includes hyper-parameter tuning, using the Random Search Forecaster from skforecast. Unlike grid search, which exhaustively searches through all possible combinations, Random Search randomly selects combinations, providing a more efficient way to find good hyperparameter values, especially when the hyperparameter space is large.

For the hydro forecast, we use 15-minute-spaced data spanning over 4 years. Our feature engineering process is tailored to capture the essential components of time series data: trend, seasonality, and cyclicality. Moving averages with a window of 35,040 (equivalent to a year's data points) are calculated to approximate the general trend. A time step feature was used to assign a unique number to each data point from 1 to 97,507. It serves as an independent variable within a linear regression framework, facilitating the estimation of 1st, 2nd, and 3rd order polynomial trends. Fourier series analysis, with a focus on identifying appropriate periodicity (initially identified as 10 but later optimized to 5 based on model performance), is employed to model seasonality more accurately. Additionally, auto-correlation analysis suggests using 8 lags as features. These incorporate data from the previous 8 time periods into the analysis to leverage the identified auto-correlation, where past values significantly influence future values in the time series. In this hybrid model, linear regression is specifically employed to capture seasonality patterns by using day-of-the-week and Fourier series components, while XGBoost regression handles the remaining features, leveraging its capability to model complex, non-linear relationships and interactions in the data. This approach is validated with metrics like MSE, RMSE, MAE, and R2, showing satisfactory performance on both the train and test sets.

D. Results

Table VI shows hourly summary 3-year forecast for DNI, wind speed and temperature for the four desalination plants used in this study. Table VII calculates the annualized 3-year average energy generation (per unit) from PV, wind and CSP using the hourly forecast weather data and RE system assumptions, which shows that energy generation from PV and CSP is highest in KBH and Alameda, and for wind is highest in Tampa and SAWS. The preliminary assessment of run-of-river or mini hydro shows sparse potential in Tampa and Alameda and mini hydro potential (100 kW-1 MW) in KBH and SAWS, due to low discharge and gage height. Future work could make a more sophisticated and optimized hydro resource characterization.

The results show that among the tested AI models, Gradient Boosting and an innovative average method of XGBoost have the best accuracy. Table V shows that the RMSE for weather forecast varied across different variables and locations. On average, the normalized RMSE (defined as $\frac{\text{RMSE}}{\text{Range of Observed Values} \times 100}$) of the XGBoost average model for Temperature, Wind Speed, Treated Water Flow and Energy Consumption was less than 10%, and for DNI approached 20% in certain cases. The code for our models used in this study can be accessed at https://github.com/anti-integral/Decarbonization-Study.

V. Conclusion

On-site RE generation allows energy independence and security and can protect against wide swings in energy prices and any potential supply disruption, thereby contributing to the overall resilience of water supply infrastructure. Specifically, a 50% RE mix could lead to a halving of the annual CO_2 output, equivalent to 16 billion gallons or 16,000 metric tons of CO_2 reduced annual emissions, akin to the benefit derived from planting 600,000 trees [23]. Our long-term energy consumption, water flow, and weather forecasting models are able to capture the trend of the input data despite its volatile trends, producing results with less than 10% normalized RMSE (up to 20% for DNI). This is a reasonable outcome while taking into consideration that these models are predicting years into the future and yet able to show patterns similar to the input data. The only areas where they don't perform are outliers, which are obviously hard to predict due to their random nature in a long forecast process.

These forecasting methods also have other practical applications, including helping researchers use these AI models to forecast climate for other projects, identify trends in how weather could change over time, and provide a framework to improve long-term weather modeling.

A. Future Research

1) Improving Weather Forecasting Models: Enhancing weather forecasting models could lead to more accurate forecasting, critical for optimal RE utilization. Future studies can focus on integrating more sophisticated ML algorithms, like deep neural networks or ensemble methods, which could offer better accuracy and reliability. Additionally, incorporating real-time data feeds and expanding the dataset to include more diverse weather scenarios could improve model robustness.

TABLE V: XGBoost Average Model Results

Forecast Quantity	Range	Alameda RMSE	SAWS RMSE	Tampa Bay RMSE	Kay Bailey RMSE
Temperature (°C)	-10.0 to 46.1	3.37	4.67	3.47	3.62
Hourly DNI ($\frac{W}{m^2}$)	0 to 1,049	189	234	199	204
Wind Speed ($\frac{m}{s}$)	0 to 35.2	1.72	2.48	2.79	2.97
Discharge (m^3/s)	RMSE/Range	2.2 (0 to 180)	4.0 (1 to 382)	1.4 (1 to 54)	4.5 (0.3 to 132)
Gage Height (m)	0 to 13	0.2	0.1	0.1	0.1
Treated Water Flow (MGD)	0 to 30	1.49	0.86	1.17	0.39
Energy Consumption (kWh)	RMSE/Range	59,000 (0 to 1,000,000)	711 (0 to 300,000)	6,751 (0 to 300,000)	10,185 (0 to 83,333)

TABLE VI: Average Weather Values, 3-Year Forecast

Site	DNI (kWh/m^2/day)	Wind Speed (m/s)	Temperature (°C)	Discharge (m^3/s)	Gage Height (m)
Tampa	5.52	5.91	23.26	0.6	5.8
SAWS	5.58	4.60	21.38	12.1	3.5
KBH	7.82	4.82	18.72	16.1	2.2
Alameda	6.77	3.88	16.11	0.6	1.7

TABLE VII: Energy Generation, 3-Year Annualized Forecast

Site	PV	Wind	CSP
Power Capacity of 1 unit (kW)	0.7	2,500	200
Tampa (kWh)	1,437	8,831,674	1,222,751
SAWS (kWh)	1,462	8,326,970	1,113,431
KBH (kWh)	2,056	5,317,925	1,560,330
Alameda (kWh)	1,752	2,129,060	1,350,030

2) Development of Optimization Models for System Sizing: Future research can develop advanced optimization models that consider various factors like the cost of installation, maintenance, the efficiency of energy conversion, and the compatibility of different RE sources and storage. Techniques such as linear programming, mixed-integer linear programming, or more advanced, AI-driven optimization methods can be employed to find the optimal mix and size of RE systems that balance cost, carbon emissions, efficiency, and reliability.

3) Conducting Sensitivity Analysis: A comprehensive sensitivity analysis (including Monte Carlo simulations) with respect to dependent variables such as weather conditions, water production rates, and energy demand is crucial. This analysis would help in understanding how changes in these parameters impact the performance of the RE systems. This knowledge is vital for designing systems that are resilient to variations in environmental conditions and operational demands.

References

[1] M. H. Saray and A. T. Haghighi, "Energy analysis in water-energy-food-carbon nexus," *Energy Nexus*, vol. 11, p. 100223, 2023.

[2] V. G. Gude and V. Fthenakis, "Energy efficiency and renewable energy utilization in desalination systems," *Progress in Energy*, vol. 2, p. 022003, 2020.

[3] Q. Li, J. Loy-Benitez, K. Nam, S. Hwangbo, J. Rashidi, and C. Yoo, "Sustainable and reliable design of reverse osmosis desalination with hybrid renewable energy systems through supply chain forecasting using recurrent neural networks," *Energy*, vol. 178, pp. 277–292, 2019.

[4] O. Sanan, J. Sperling, D. Greene, and R. Greer, "Towards data-driven methods for decarbonizing reverse osmosis desalination," in *2023 IEEE MIT Undergraduate Research Technology Conference*. IEEE, 2023.

[5] A. Tal, "Addressing desalination's carbon footprint: the israeli experience," *Water*, vol. 10, no. 2, p. 197, 2018.

[6] M. K. Shahzad, A. Zahid, T. ur Rashid, M. A. Rehan, M. Ali, and M. Ahmad, "Techno-economic feasibility analysis of a solar-biomass off grid system for the electrification of remote rural areas in pakistan using homer software," *Renewable energy*, vol. 106, pp. 264–273, 2017.

[7] E. T. Sayed, A. Olabi, K. Elsaid, M. Al Radi, R. Alqadi, and M. A. Abdelkareem, "Recent progress in renewable energy based-desalination in the middle east and north africa mena region," *Journal of Advanced Research*, vol. 48, pp. 125–156, 2023.

[8] T. Ayodele and A. Ogunjuyigbe, "Mitigation of wind power intermittency: Storage technology approach," *Renewable and Sustainable Energy Reviews*, vol. 44, pp. 447–456, 2015.

[9] J. Leijon and C. Boström, "Freshwater production from the motion of ocean waves–a review," *Desalination*, vol. 435, pp. 161–171, 2018.

[10] R. J. Hyndman and G. Athanasopoulos, *Forecasting: principles and practice*. OTexts, 2018.

[11] J. A. Carta and P. Cabrera, "Optimal sizing of stand-alone wind-powered seawater reverse osmosis plants without use of massive energy storage," *Applied Energy*, vol. 304, p. 117888, 2021.

[12] F. Rodríguez, A. Fleetwood, A. Galarza, and L. Fontán, "Predicting solar energy generation through artificial neural networks using weather forecasts for microgrid control," *Renewable energy*, vol. 126, 2018.

[13] L. Chen, B. Han, X. Wang, J. Zhao, W. Yang, and Z. Yang, "Machine learning methods in weather and climate applications: A survey," *Applied Sciences*, vol. 13, no. 21, p. 12019, 2023.

[14] A. G. Salman and B. Kanigoro, "Visibility forecasting using autoregressive integrated moving average (arima) models," *Procedia Computer Science*, vol. 179, pp. 252–259, 2021.

[15] Y. Liu, M. C. Roberts, and R. Sioshansi, "A vector autoregression weather model for electricity supply and demand modeling," *Journal of Modern Power Systems and Clean Energy*, vol. 6, no. 4, pp. 763–776, 2018.

[16] A. M. Abdelshafy, H. Hassan, and J. Jurasz, "Optimal design of a grid-connected desalination plant powered by renewable energy resources using a hybrid pso–gwo approach," *Energy conversion and management*, vol. 173, pp. 331–347, 2018.

[17] National Renewable Energy Laboratory. (2023) Annual technology baseline. [Online]. Available: https://atb.nrel.gov/electricity

[18] "Renewable power generation costs in 2022," in *irena.org*. International Renewable Energy Agency (IRENA), 2022.

[19] M. Sengupta, Y. Xie, A. Lopez, A. Habte, G. Maclaurin, and J. Shelby, "The national solar radiation data base (nsrdb)," *Renewable and sustainable energy reviews*, vol. 89, pp. 51–60, 2018.

[20] B. Hadjerioua, S. Kao, M. Sale, Y. Wei, S. SanthanaVannan, H. Shanafield III, D. Kaiser, R. Devarakonda, C. Odeh, G. Palanisamy *et al.*, "National hydropower asset assessment project," *US Department of Energy, Oak Ridge National Laboratory*, 2011.

[21] D. Spera and T. Richards, "Modified power law equations for vertical wind profiles," in *Proceedings of the Conference and Workshop on Wind Energy Characteristics and Wind Energy Siting; 19-21 June 1979; Portland, Oregon (USA)*, 1979, pp. 47–58.

[22] S. J. Taylor and B. Letham, "Forecasting at scale," *The American Statistician*, vol. 72, no. 1, pp. 37–45, 2018.

[23] A. Moseman, C. Harvey, and C. Terrer, "How many new trees would we need to offset our carbon emissions?" in *Ask MIT Climate*. https://climate.mit.edu/ask-mit/how-many-new-trees-would-we-need-offset-our-carbon-emissions, 2024.