Gabriel Mendonça de Paiva

Machine Learning Methods Applied to Intraday Solar Forecasting for Power System Operation

Goiania - GO

2021

UNIVERSIDADE FEDERAL DE GOIÁS
ESCOLA DE ENGENHARIA ELÉTRICA, MECÂNICA E DE COMPUTAÇÃO

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO (TECA) PARA DISPONIBILIZAR VERSÕES ELETRÔNICAS DE TESES

E DISSERTAÇÕES NA BIBLIOTECA DIGITAL DA UFG

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio da Biblioteca Digital de Teses e Dissertações (BDTD/UFG), regulamentada pela Resolução CEPEC nº 832/2007, sem ressarcimento dos direitos autorais, de acordo com a Lei 9.610/98, o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou download, a título de divulgação da produção científica brasileira, a partir desta data.

O conteúdo das Teses e Dissertações disponibilizado na BDTD/UFG é de responsabilidade exclusiva do autor. Ao encaminhar o produto final, o autor(a) e o(a) orientador(a) firmam o compromisso de que o trabalho não contém nenhuma violação de quaisquer direitos autorais ou outro direito de terceiros.

1. Identificação do material bibliográfico

[ ] Dissertação [X] Tese

2. Nome completo do autor

Gabriel Mendonça de Paiva

3. Título do trabalho

"Machine Learning Methods Applied to Intraday Solar Forecasting for Power System Operation"

4. Informações de acesso ao documento (este campo deve ser preenchido pelo orientador)

Concorda com a liberação total do documento [X] SIM [ ] NÃO[1]

[1] Neste caso o documento será embargado por até um ano a partir da data de defesa. Após esse período, a possível disponibilização ocorrerá apenas mediante:
a) consulta ao(à) autor(a) e ao(à) orientador(a);
b) novo Termo de Ciência e de Autorização (TECA) assinado e inserido no arquivo da tese ou dissertação.
O documento não será disponibilizado durante o período de embargo.
Casos de embargo:
- Solicitação de registro de patente;
- Submissão de artigo em revista científica;
- Publicação como capítulo de livro;
- Publicação da dissertação/tese em livro.

Obs. Este termo deverá ser assinado no SEI pelo orientador e pelo autor.

Documento assinado eletronicamente por **Bernardo Pinheiro De Alvarenga**, **Professor do Magistério Superior**, em 25/03/2021, às 13:58, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **GABRIEL MENDONÇA DE PAIVA**, **Discente**, em

Processo:
23070.014606/2021-37

Documento:
1950721

Referência: Processo nº 23070.014606/2021-37 SEI nº 1950721

Gabriel Mendonça de Paiva

Machine Learning Methods Applied to Intraday Solar Forecasting for Power System Operation

Thesis presented to the Graduate Program (*Stricto Sensu*) in Electrical and Computer Engineering as a partial requirement for obtaining the doctoral degree in Electrical and Computer Engineering.

Knowledge field: Electrical Engineering / Computer Engineering

Federal University of Goias – UFG

Electrical Mechanic and Computer Engineering School - EMC

Graduate Program in Electrical and Computer Engineering

Supervisor: Bernardo Pinheiro de Alvarenga, PhD

Co-supervisor: Enes Gonçalves Marra, PhD

Goiania - GO

2021

Ficha de identificação da obra elaborada pelo autor, através do
Programa de Geração Automática do Sistema de Bibliotecas da UFG.

Paiva, Gabriel Mendonça de
Machine Learning Methods Applied to Intraday Solar Forecasting for Power System Operation [manuscrito] / Gabriel Mendonça de Paiva. - 2021.
cii, 102 f.: il.

Orientador: Prof. Dr. Bernardo Pinheiro de Alvarenga; co orientador Dr. Enes Gonçalves Marra.
Tese (Doutorado) - Universidade Federal de Goiás, Escola de Engenharia Elétrica, Mecânica e de Computação (EMC), Programa de Pós-Graduação em Engenharia Elétrica e de Computação, Goiânia, 2021.
Bibliografia. Apêndice.
Inclui siglas, mapas, fotografias, abreviaturas, símbolos, gráfico, tabelas, algoritmos, lista de figuras, lista de tabelas.

1. Artificial neural networks. 2. Intraday solar forecasting. 3. Multigene genetic programming. 4. Power system operation. 5. PV power forecasting. I. Alvarenga, Bernardo Pinheiro de, orient. II. Título.

CDU 621.3

UNIVERSIDADE FEDERAL DE GOIÁS

ESCOLA DE ENGENHARIA ELÉTRICA, MECÂNICA E DE COMPUTAÇÃO

ATA DE DEFESA DE TESE

Ata Nº **03** da sessão de Defesa de Tese de Gabriel Mendonça de Paiva que confere o título de Doutor em **Engenharia Elétrica e de Computação**, na área de concentração em **Engenharia Elétrica.**

Aos **cinco dias do mês de março de dois mil e vinte e um**, a partir das **09h00min.**, realizou-se a sessão pública de Defesa de Tese intitulada **"Machine Learning Methods Applied to Intraday Solar Forecasting for Power System Operation**". Os trabalhos foram instalados pelo Orientador, Professor Doutor **Bernardo Pinheiro de Alvarenga (EMC/UFG)** com a participação dos demais membros da Banca Examinadora: Professor Doutor **Marco Mussetta (POLIMI/IT)** membro titular externo, Professor Doutor **Enes Gonçalves Marra (EMC/UFG)** membro titular externo, Professor Doutor **Sergio Pires Pimentel (EMC/UFG)** membro titular externo e Professor Doutor **Leonardo da Cunha Brito (EMC/UFG)** membro titular interno**, cujas participações ocorreram através de videoconferência**. Durante a argüição os membros da banca **não fizeram** sugestão de alteração do título do **trabalho**. A Banca Examinadora reuniu-se em sessão secreta a fim de concluir o julgamento da Tese tendo sido o candidato **aprovado** pelos seus membros. Proclamados os resultados pelo Professor Doutor Bernardo Pinheiro de Alvarenga, Presidente da Banca Examinadora, foram encerrados os trabalhos e, para constar, lavrou-se a presente ata que é assinada pelos Membros da Banca Examinadora, aos cinco dias do mês de março de dois mil e vinte e um.

TÍTULO SUGERIDO PELA BANCA

---

sei! assinatura eletrônica

Documento assinado eletronicamente por **Bernardo Pinheiro De Alvarenga**, **Professor do Magistério Superior**, em 05/03/2021, às 11:02, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

---

sei! assinatura eletrônica

Documento assinado eletronicamente por **Sergio Pires Pimentel**, **Professor do Magistério Superior**, em 05/03/2021, às 11:03, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

---

sei! assinatura eletrônica

Documento assinado eletronicamente por **MARCO MUSSETTA**, **Usuário Externo**, em 05/03/2021, às 11:04, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

---

sei! assinatura eletrônica

Documento assinado eletronicamente por **Enes Goncalves Marra**, **Professor do Magistério Superior**, em 05/03/2021, às 11:04, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Leonardo Da Cunha Brito**, **Professor do Magistério Superior**, em 05/03/2021, às 11:04, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **GABRIEL MENDONÇA DE PAIVA**, **Discente**, em 05/03/2021, às 11:26, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **1898832** e o código CRC **B06DB192**.

Referência: Processo nº 23070.007479/2021-10 SEI nº 1898832

I dedicate this thesis to my parents, who supported my studies at all stages. They lived, believed and dreamed the good and hard moments throughout the research in Brazil and in Italy.

Acknowledgements

I am thankful, firstly, for my supervisors, Bernardo Alvarenga and Enes Gonçalves Marra, who defined and participated in the formulation and development of this research since the first stage.

Secondly, I thank for my parents, brothers and friends, who supported my career as a researcher.

I thank the Federal University of Goias and the Electrical Mechanic and Computer Engineering School for providing access and allow for scientific usage of the research data. I am also thankful for EMC–UFG for being the main part of my career development as an engineer and as a researcher, where I studied for 12 years. Studying, working and making friends in the University has been a great experience.

I thank Espora Energetica S/A, proponent company of the R&D project related to this research, and the respective cooperative companies, Transenergia Sao Paulo S/A, Transenergia Renovavel S/A, and Caldas Novas Transmissao S/A.

I thank Politecnico di Milano, in name of professors Marco Mussetta and Sonia Leva, for contributing to this research and making part of my career development. I thank Politecnico di Milano also for providing access and allow for scientific usage of their meteorological station data.

I thank the National Oceanic and Atmospheric Administration (NOAA) for providing access and allow for scientific usage of their meteorological stations data.

"The ideal is still the soul of all achievements."
Getulio Vargas

Resumo

Prever o recurso solar é uma ferramenta essencial para sua integração com a rede elétrica. Esta tese foca em previsão solar intra-diária, com uma análise robusta de previsão de irradiância testada em múltiplas localidades e uma proposta de implementação de previsão de potência fotovoltaica (FV). Dois algoritmos de aprendizagem de máquinas são avaliados para previsão intra-diária da irradiância solar: programação genética multigene (PGMG) e redes neurais artificiais do tipo multilayer perceptron (MLP). PGMG é um algoritmo evolucionário e um método tipo "caixa branca" e é uma nova técnica na área. Os algoritmos de aprendizagem de máquinas também são comparados com um modelo de persistência inteligente (*smart persistence*) para prever a irradiância solar com dados de seis localidades. Os horizontes de previsão considerados são 15–120 minutos à frente. Os resultados das simulações mostram um aprimoramento consistente das previsões quando variáveis climáticas exógenas são adicionadas como entrada aos modelos, sendo 5.68% o aprimoramento pelo cálculo de erro médio absoluto (MAE) e 3.41% o aprimoramento pelo cálculo de raiz do erro quadrático médio ($RMSE$). Os resultados também mostram que localidade, horizonte de previsão e métrica de erro escolhida influenciam a dominância de acurácia dos modelos. Dois modelos de irradiância de céu claro foram implementados, mas os resultados indicam para uma baixa influência dos modelos na acurácia de previsão para previsões multivariadas por aprendizagem de máquinas. Em uma perspectiva genérica, PGMG apresentou resultados mais precisos e robustos que MLP em previsões individuais, provendo soluções mais rápidas. Entretanto, MLP apresentou mais precisão em previsões do tipo *ensemble*, porém estas apresentam também maior complexidade e maior custo computacional. A implementação de previsão de potência FV mostrou resultados consistentes, aprimorando valores de $RMSE$ de previsões de persistência em 9.79%–23.75% para horizontes de 15–120 minutos.

Palavras-chave: Redes neurais artificiais; previsão solar intra-diária; programação genética multigene; operação de sistema elétrico de potência; previsão de potência FV; previsão de curto prazo.

Abstract

Forecasting solar resources is an essential tool for its integration into electrical utility grids. This thesis focuses on intraday solar forecasting, with a robust analysis of irradiance forecasting tested on multiple sites and a proposed PV power forecast implementation. Two machine learning (ML) algorithms for intraday solar irradiance forecasting were evaluated: multigene genetic programming (MGGP) and the multilayer perceptron (MLP) artificial neural network (ANN). MGGP is an evolutionary algorithm white-box method and is a novel approach in the field. The ML algorithms are also compared with a smart persistence model to forecast irradiance for databases from six locations. The forecast horizons under consideration are 15–120 minutes ahead. The results show a consistent improvement of MAE and RMSE values by adding exogenous weather variables as input to the ML algorithms of 5.68% and 3.41%, respectively. The results also show that location, forecast horizon and error metric definition affect model accuracy dominance. Two clear sky models were implemented, but results indicate a low influence of them in multivariate ML forecast accuracy. MGGP presented more accurate and robust results in single prediction cases in a general perspective, providing faster solutions. In contrast, ANN presented more accurate ensemble forecasting results, although it presented higher complexity and required additional computational effort. The implementation of the PV power forecasting model has shown consistent results, improving RMSE values from a k_{PV} index smart persistence by 9.79%–23.75% for horizons of 15–120 minutes.

Keywords: Artificial neural networks; Intraday solar forecasting; Multigene genetic programming; Power system operation; PV power forecasting; Short-term forecasting.

List of Figures

Figure 2.1 – Extraterrestrial irradiance (G_o) in a horizontal plane located in Goiania at an yearly point of view (top) and daily point of view (bottom). . . . 27
Figure 2.2 – Global horizontal irradiance (G) curves from EMC-UFG Weather Station at an yearly point of view (top) and daily point of view (bottom) - year of reference: 2016. 27
Figure 2.3 – Spatial resolution and forecast horizon domains of different solar forecasting methods by the classification of Inman, Pedro and Coimbra (2013). 29
Figure 2.4 – Spatial resolution and forecast horizon domains of different solar forecasting methods by the classification of Diagne et al. (2013). 30
Figure 2.5 – Deterministic 5 parameters PV equivalent circuit for physical PV power forecasting applied by Ogliari et al. (2017). 34
Figure 2.6 – An example of the cloud indexing procedure. a) Reference ground albedo image obtained by taking the minimum pixel intensities for each pixel location. b) Original image taken on 11/23/2011. c) Cloud indexed image after applying normalization. The two dots identify the forecasting locations. North-most Davis. South-most Merced. 35
Figure 2.7 – Relations between sky images, cloud detection and irradiance measurements in the methodology of Pedro et al. (2018). 36
Figure 2.8 – Improvement of Autoregressive (AR) forecasts for 1-6h forecast horizons by application of a spatio-temporal model. 38
Figure 2.9 – Structure of a single neuron with inputs $p_1, ..., p_R$, weights $w_{1,1}, ..., w_{1,R}$, a bias b and an output a. 40
Figure 2.10–Structure of a multiple-layer neural network. 40
Figure 2.11–Example of a separable problem in a bi-dimensional space by Cortes and Vapnik (1995). Support vectors are marked with grey squares and define the margin of largest separation between two classes. 41
Figure 2.12–Analysis of day-ahead forecast percentage errors in a sort distribution by Leva et al. (2017). 46
Figure 3.1 – Flowchart of the intraday solar irradiance forecasting methodology. . . 48
Figure 3.2 – Locations of the meteorological stations under analysis presented in a world map. 49
Figure 3.3 – Global Horizontal Irradiance (GHI) curves of Goiania and Milan complete datasets in 15-min time windows. 51
Figure 3.4 – Global Horizontal Irradiance (GHI) curves of Desert Rock and Pennsylvania SU complete datasets in 15-min time windows. 52

Figure 3.5 – Global Horizontal Irradiance (GHI) curves of Bondville and Sioux Falls complete datasets in 15-min time windows. 52
Figure 3.6 – Global Horizontal Irradiance (GHI) curves for 2 days in Milan dataset presenting raw 1-min data (blue curve) and processed 15-min time windows data (red curve). 53
Figure 3.7 – Relative humidity graph of a specific year with an outlier (top) and after being pre-processed by the data quality control algorithm (bottom). 54
Figure 3.8 – Linke turbidity daily fitness (blue lines) and monthly averaged values (red dots) for each location. 56
Figure 3.9 – Ineichen clear sky index curve for the complete dataset of Goiania (top) and for 5 winter and summer days (bottom) (15-min time windows data). 57
Figure 3.10 – Clear sky index curves of complete data sets. From top to bottom: Milan, Desert Rock and Pennsylvania SU. 58
Figure 3.11 – Clear sky index curves of complete data sets. From top to bottom: Bondville and Sioux Falls. 59
Figure 3.12 – Data relations of the MISO system methodology for intraday solar irradiance forecasting. 60
Figure 3.13 – Example of a multigene symbolic regression (SR) model presented in a tree structure (PAIVA et al., 2020). 62
Figure 3.14 – Fitness of best GP solution s^* measured by k_t^* RMSE for training and validation datasets (PAIVA et al., 2020). 64
Figure 3.15 – Data relations in iterative forecasts. 65
Figure 4.1 – Influence of maximum number of genes on the fitness of best solutions, evaluated for the Goiania validation dataset (PAIVA et al., 2020). . . . 68
Figure 4.2 – Influence of tournament size (κ), mutation rate (p_m) and elitism rate (elit) on the accuracy and robustness (RMSE standard deviation) of the validation dataset from the Goiania site (PAIVA et al., 2020). 68
Figure 4.3 – Improvements (%) of multivariate forecasting using GP according to mean absolute error (MAE) (dark red bars) and according to RMSE (dark blue bars), and improvements (%) of multivariate forecasting using an artificial neural network (ANN) according to MAE (orange bars) and according to RMSE (light blue bars) (PAIVA et al., 2020). 69
Figure 4.4 – Model accuracy dominance by location and forecast horizon in single forecasts. GP/ANN indicates cases in which accuracy dominance differs from the error metric evaluated (PAIVA et al., 2020). 70
Figure 4.5 – Accuracy of persistence, GP and ANN according to RMSE (**left**) and MAE (**right**) for Goiania, showing the dominance of ANN (PAIVA et al., 2020). 71

Figure 4.6 – Accuracy of persistence, GP and ANN according to RMSE (**left**) and MAE (**right**) for Desert Rock, showing the dominance of GP (PAIVA et al., 2020). 71
Figure 4.7 – Model accuracy dominance by location and forecast horizon in ensemble forecasts. GP/ANN indicates cases in which accuracy dominance differs from the error metric evaluated (PAIVA et al., 2020). 72
Figure 4.8 – General accuracies of GP, ANN, GP ensemble and ANN ensemble for all sites according to RMSE values (left-hand graphs) and MAE values (right-hand graphs) (PAIVA et al., 2020). 73
Figure 4.9 – Comparison of general robustness of GP and ANN single forecasts according to MAE and RMSE (PAIVA et al., 2020). 73
Figure 4.10–Individual errors plot from MGGP forecasts for Goiania test dataset for forecast horizons of 15-min (top), 60-min (mid) and 120-min (bottom). 76
Figure 4.11–Monthly RMSE plots from $MGGP_{ens}$ (blue bars) and ANN_{ens} (red bars) forecasts for Goiania test dataset for forecast horizons of 90-min (top), 105-min (mid) and 120-min (bottom). 76
Figure 5.1 – Google earth view of the EMC–UFG grid connected PV system. 79
Figure 5.2 – Virtual 1 kWp PV plant output power (top) and k_{PV} distribution (bottom). 80
Figure 5.3 – Data relations of input and output vectors for the case of PV output power forecast model developed. 80
Figure 5.4 – Table of comparison of the forecast results from ANN ensemble versus a k_{PV} persistence from Paiva et al. (2019) (best results presented in bold) . 81

List of Tables

Table 2.1 – Some of the main review articles in solar forecasting literature. 28

Table 2.2 – Characteristics of data relations of point forecast studies in literature based on meteorological stations for irradiance forecasting (PEDRO; COIMBRA, 2015b; NOBRE et al., 2016) and on PV power meters (RANA; KOPRINSKA; AGELIDIS, 2016). $\Delta\tau$ is the forecast horizon, Δt_{in} is the time resolution of input data, Δt_{out} is the time resolution of output data. MAE and s are error metrics defined in 2.5. 32

Table 2.3 – Table of comparison of single machine learning methods for solar forecasting by Voyant et al. (2017). 44

Table 2.4 – Table of comparison of ensemble machine learning methods for solar forecasting by Voyant et al. (2017) (LR = Linear Regression). 44

Table 3.1 – Description of the equipment used at the Federal University of Goias (UFG) weather station, the parameters they measure and their accuracy and range of operation. 50

Table 3.2 – Statistics of Global Horizontal Irradiance (GHI) for each database: total yearly GHI in kWh/m^2, average GHI in 15-min time windows μ_{15} and standard deviation of GHI in 15-min time windows σ_{15}. 53

Table 3.3 – Data statistics of training, validation and test datasets for each location: N_{samp} is the number of samples of each dataset, μ is the average Ineichen k_t^* and σ is the standard deviation of k_t^*. 57

Table 3.4 – Summary of genetic programming (GP) simulation parameters. 63

Table 4.1 – Generalized accuracies for Haurwitz k_t^* and Ineichen k_t^* forecasts (PAIVA et al., 2020). 73

Table 4.2 – Variables selected by GP regression models according to the forecast horizon for Goiania (PAIVA et al., 2020). 74

Table 4.3 – Comparison of state-of-the-art methods applied to intraday solar irradiance forecasting for Surface Radiation Network (SURFRAD) weather stations (best values in bold) (PAIVA et al., 2020). 75

Table 4.4 – Comparison of training time required for each machine learning (ML) method, evaluated for the Goiania dataset (best values in bold) (PAIVA et al., 2020). 75

Table .1 – Forecast errors for Goiania (best values in bold). *Persist*: persistence. . 96

Table .2 – Forecast errors for Milan (best values in bold). 97

Table .3 – Forecast errors for Desert Rock (best values in bold). 98

Table .4 – Forecast errors for Pennsylvania State University (best values in bold). . 99

Table .5 – Forecast errors for Bondville (best values in bold). 100

Table .6 – Forecast errors for Sioux Falls (best values in bold). 101

List of abbreviations and acronyms

- AI: artificial intelligence
- AM: air mass
- ANN: artificial neural network
- AR: autoregressive
- ARIMA: autoregressive integrated moving average
- ARMA: autoregressive moving average
- ARMAX: autoregressive moving average with exogenous inputs
- ARX: autoregressive with exogenous inputs
- DC: direct current
- DNI: direct normal irradiance
- EA: evolutionary algorithm
- ECMWF: European Centre for Medium-Range Weather Forecasts
- EMAE: envelope-weighted mean absolute error
- EMC: Escola de Engenharia Elétrica Mecânica e de Computação (School of Electrical Mechanic and Computer Engineering)
- ERC: ephemeral random constants
- FFNN: feedforward neural network
- GA: genetic algorithm
- GB: gradient boosting
- GBR: gradient boosted regression
- GBRT: gradient boosted regression trees
- GFS: global forecast system
- GHI: global horizontal irradiance
- GP: genetic programming

- IEA: International Energy Agency
- k-NN: k-nearest neighbors
- LR: linear regression
- MAE: mean absolute error
- MAPE: mean absolute percentage error
- MBE: mean bias error
- MGGP: multigene genetic programming
- MILP: mixed-integer linear programming
- MIMO: multi-input multi-output
- MISO: multi-input single output
- ML: machine learning
- MLP: multilayer perceptron
- nMAE: normalized mean absolute error
- NNens: neural networks ensemble
- NOAA: National Oceanic and Atmospheric Administration
- nRMSE: normalized root mean square error
- NWP: numerical weather prediction
- OMAE: objective mean absolute error
- PCA: principal component analysis
- PHANN: physical hybrid artificial neural network
- PoliMi: Politecnico di Milano
- PSU: Pennsylvania State University
- PV: photovoltaic
- RBF: radial basis function
- RES: renewable energy sources
- RF: random forest

- RMSE: root mean square error
- RNN: recurrent neural network
- RT: regression trees
- SARIMA: seasonal autoregressive moving average
- SURFRAD: Surface Radiation Network
- SVM: support vector machine
- SVR: support vector regression
- SR: symbolic regression
- UFG: Universidade Federal de Goiás (Federal University of Goias)
- UAE: United Arab Emirates
- USA: United States of America
- WRF: Weather Research and Forecasting

List of symbols

- $elit$ elitism rate
- G global horizontal irradiance
- G_b beam horizontal irradiance
- G_{clr} clear sky irradiance
- $G_{clr,T}$ transposed clear sky irradiance
- G_d diffuse horizontal irradiance
- G_o extraterristrial irradiance
- G_{sc} solar constant
- h solar elevation angle
- H_r relative humidity
- k_t clearness index
- k_t^* clear sky index
- k_{PV} photovoltaic clear sky index
- p_a atmospheric pressure
- p_c crossover probability
- p_m mutation probability
- P_o PV output power
- p_r reproduction probability
- s forecast skill
- T_a ambient temperature
- t_s time difference in respect to sunrise
- β surface tilt angle
- δ declination angle
- γ_s surface azimuth angle

- κ tournament size
- ϕ local latitude
- θ_z solar zenith angle
- ω hour angle
- ω_s solar time angle

Published Studies

Journals:

- PAIVA G. M.; PIMENTEL, S. P.; ALVARENGA, B. P.; MARRA, E. G.; MUSSETTA, M.; LEVA, S.; Multiple sites intraday solar irradiance forecasting by machine learning algorithms: MGGP and MLP Neural Networks. *Energies*, v. 13, p. 3005, 2020. (PAIVA et al., 2020)
- PAIVA G. M.; PIMENTEL, S. P.; ALVARENGA, B. P.; MARRA, E. G.; REGRESSÃO SIMBÓLICA APLICADA NA PREVISÃO DE IRRADIÂNCIA SOLAR INTRA-DIÁRIA NA CIDADE DE GOIÂNIA (BRASIL). *Revista Brasileira de Energia Solar*, v. 9, n. 2, 2018.(PAIVA SéRGIO PIRES PIMENTEL, 2018)

Conferences:

- PAIVA G. M.; PIMENTEL, S. P.; MARRA, E. G.; ALVARENGA, B. P.; MUSSETTA, M.; LEVA, S.; Intra-day forecasting of building-integrated PV systems for power systems operation using ANN ensemble. *IEEE Milan Powertech*, 2019, Milan. p. 1-5. (Paiva et al., 2019)
- PAIVA G. M.; PIMENTEL, S. P.; LEVA, S.; MUSSETTA, M.; Assessment of Exogenous Variables on Intra-Day Solar Irradiance Forecasting Models. *2018 IEEE International Conference on Environment and Electrical Engineering and 2018 IEEE Industrial and Commercial Power Systems Europe*, 2018, Palermo. p. 1-6. (de Paiva et al., 2018a)
- PAIVA G. M.; PIMENTEL, S. P.; LEVA, S.; MUSSETTA, M.; Intelligent Approach to Improve Genetic Programming Based Intra-Day Solar Forecasting Models. *2018 IEEE Congress on Evolutionary Computation (CEC)*, 2018, Rio de Janeiro. p. 1-8. (de Paiva et al., 2018b)
- PAIVA G. M.; PIMENTEL, S. P.; MARRA, E. G.; ALVARENGA, B. P.; REGRESSÃO SIMBÓLICA APLICADA NA PREVISÃO DE IRRADIÂNCIA SOLAR INTRA-DIÁRIA NA CIDADE DE GOIÂNIA (BRASIL). *VII Congresso Brasileiro de Energia Solar*, Anais do VII Congresso Brasileiro de Energia Solar, 2018.
- PAIVA G. M.; PIMENTEL, S. P.; ALVARENGA, B. P.; MARRA, E. G.; Programação Genética Multigene aplicada a previsão de curto prazo de Irradiância Global Horizontal em Goiânia. *14° Congresso de Ensino Pesquisa e Extensão*, 2017, Goiânia. Anais 14° CONPEEX, 2017.

Contents

1 Introduction **23**
- 1.1 Main objective 24
- 1.2 Secondary objectives 25
- 1.3 Summary 25

2 Solar Forecasting Research **26**

This chapter brings a comprehensive literature review about the solar forecasting research field and presents the main aspects that have driven methodologies in the field.

- 2.1 Solar Energy Variability 26
- 2.2 Solar Forecasting Reviews 28
- 2.3 Types of data acquisition systems 30
  - 2.3.1 Meteorological stations and power meters 30
  - 2.3.2 Numerical Weather Prediction 33
  - 2.3.3 Satellite data acquisition 34
  - 2.3.4 Ground based sky imagers 35
- 2.4 Statistical methods for solar forecasting 36
  - 2.4.1 Persistence and Smart Persistence 36
  - 2.4.2 Autoregressive method 37
  - 2.4.3 Autoregressive integrated moving average (ARIMA) models 38
  - 2.4.4 Artificial neural networks 39
  - 2.4.5 Support Vector Machines 41
  - 2.4.6 k-Nearest Neighbors 42
  - 2.4.7 Random Forest 42
  - 2.4.8 Gradient Boosted Regression Trees 43
  - 2.4.9 Comparison of forecasting methods 43
- 2.5 Error metrics for solar forecasting 45
- 2.6 Summary 47

3 Intraday Solar Irradiance Forecasting using Machine Learning Algorithms . **48**

This chapter describes the proposed methodology to analyze machine learning algorithms in multiple sites.

- 3.1 Data Collection 49
  - 3.1.1 Goiania, Brazil 49
  - 3.1.2 Milan, Italy 50
  - 3.1.3 SURFRAD stations, USA 50
- 3.2 Data Pre-processing 51
  - 3.2.1 Visualization of Irradiance Curves 51

3.2.2 Data Quality Control . . . 53
3.2.3 Normalization and useless data removal . . . 54
3.2.4 Clear sky index curves and statistics . . . 56
3.3 Data relations . . . 59
3.4 Multigene genetic programming . . . 61
3.5 Multilayer perceptron (MLP) Artificial Neural Networks . . . 64
3.6 Ensemble Forecasts . . . 64
3.7 Iterative Forecasts . . . 65
3.8 Persistence . . . 65
3.9 Summary . . . 66
4 Case study I: Comparison of MGGP and MLP Neural Networks for intraday irradiance forecasting at multiple sites . . . 67
The first case results are presented in this chapter, addressing intraday solar irradiance forecasting based on data from six meteorological stations situated in Brazil, Italy, and the USA.
4.1 MGGP tuning . . . 67
4.2 Assessment of exogenous input variables . . . 69
4.3 Specific Results . . . 70
4.4 Generic Results . . . 72
4.5 Regression Functions . . . 73
4.6 Comparison with the State-of-the-Art . . . 74
4.7 Machine Learning Algorithm Training Speed . . . 74
4.8 Analysis of individual and monthly errors . . . 75
4.9 Summary . . . 77
5 Case study II: Implementation and validation of multivariate intraday PV output power forecasting . . . 78
This chapter presents results from tests of the second case covering intraday forecasting of PV output power from the EMC–UFG grid-connected PV system.
5.1 Combining meteorological station and PV power measurements . . . 78
5.2 Data processing . . . 78
5.3 Input-output scheme . . . 80
5.4 Forecasting results . . . 81
5.5 Summary . . . 81
6 Conclusions . . . 82

Bibliography . . . 84

1 Introduction

Power systems are traditionally dependent on fossil fuels, which are energy sources whose power plants emit a high amount of carbon dioxide and other carbon compounds (high carbon footprints) (MESSAGIE et al., 2014; WANG et al., 2018). In the USA, electricity from coal, natural gas and petroleum accounted for 69.2% and 62.9% of total electricity production in 2009 and 2019, respectively (EIA, 2020). In Europe, electricity generated from fossil fuels accounted for 51.3% and 40% in 2009 and 2018, respectively (EEA, 2012; EUROSTAT, 2020). In Brazil, electricity generated from fossil fuels accounted for 15.3% and 14.2% in 2008 and 2018, respectively (EPE, 2009; EPE, 2019).

Solar photovoltaic (PV) and wind energy have been the new bet for reducing carbon emissions from traditional power systems, since carbon emissions are known as greenhouse gases and are considered the leading cause of climate change (JACOBSON, 2010). In the USA, solar and wind accounted for 1.9% and 9.0% of total electricity production in 2009 and 2019, respectively (EIA, 2020). In Europe, they increased from 4.5% in 2009 to 15% in 2018 (EUROSTAT, 2020). In Brazil, they increased from nearly 0% in 2008 to 8.1% in 2018 (EPE, 2009; EPE, 2019).

In contrast to the climate benefit, the increase of intermittent renewable energy sources (RES) penetration (solar and wind) into power systems has created a challenge to systems operation (Ela; O'Malley, 2012). In order to avoid stability issues, power systems operators need to keep the balance between production and demand permanently (Kundur et al., 2004). Traditional power plants scheduling has relied on predictions of variable loads (Barcelo; Rastgoufard, 1997). Now, scheduling and system operation also need to consider forecasting of variable power production from solar and wind energy (Bakirtzis; Biskas, 2017).

Examples of stochastic and deterministic unified unit commitment and economic dispatch models for power systems operation that are based on intermittent RES forecasting are presented in Bakirtzis and Biskas (2017). The optimization models were formulated as mixed-integer linear programming (MILP) problems that own an objective function that minimizes the sum of a) unit marginal cost function, b) start-up and shut down costs, c) system reserve costs and d) cost of load and wind power curtailment (same concepts can be applied for integration of solar power into the grid). Load and RES forecasts were inputs of power balance equations that were inserted in the models as constraints. In order to support operation reliably, the power system needs to be optimized separately for each subsequent future time-window, usually in 15-minute resolution for short horizons and in 1-hour resolution for longer horizons.

The International Energy Agency (IEA) classifies solar energy forecasts in terms of the horizons evaluated as: intraday forecasting, when the target of forecasts is from a few minutes up to 6h ahead; and day-ahead forecasting, when predictions are performed for the next day (PELLAND et al., 2013). The IEA report also suggests that statistical techniques such as time-series machine learning provide good performances in the intraday context, while physical models based on numerical weather prediction (NWP) provide good performances in the day-ahead context.

For several reasons, research is needed to develop new studies and solar forecasting methods for better power systems operation. Firstly, literature review shows that the measured values of forecasting errors from different studies are generally not negligible. For example, in Aguiar et al. (2016), Root Mean Square Error (RMSE) values in a range of 83-120 W/m^2 were obtained for solar irradiance forecasts from 1h to 6h horizons at Pozo Izquierdo (Spain). At Las Palmas (Spain), the RMSE values obtained were in a range of 104-147 W/m^2 evaluated at the same horizons. In Kallio-Myers et al. (2020), RMSE values in a range of 94-184 W/m^2 were obtained for solar irradiance forecasts from 1h to 4h horizons at some sites in Finland.

Another reason for the need of solar forecasting development is that literature review does not provide a decisive method that results in best accuracy for every case study. A review of solar forecasting research indicates that machine learning is probably the most adopted methodology in the field (YANG et al., 2018), mainly due to suitable performance of algorithms when enough data is available. Specific review of machine learning methods for solar radiation forecasting concluded that ranking these methods in literature is a not a simple task (VOYANT et al., 2017), as numerous factors influence comparison among different studies and their results, namely differences in locations and datasets analyzed; processing strategies; data acquisition systems; time resolution and forecast horizons; performance indicators evaluated, etc.

Most studies rely on single location analysis, which does not bring potential generalization of results, as the solar resource presents stochastic nature in spatial and temporal levels. Sperati et al. (2015) concluded that more methods need to be evaluated at more locations to improve the solar forecasting field.

1.1 Main objective

Based on this initial background, the objective of this research is to evaluate and compare novel and traditional machine learning (ML) methods for intraday solar irradiance forecasting at multiple sites: Multigene Genetic programming (MGGP) (novel) and multilayer perceptron (MLP) Neural Networks (traditional).

The datasets analyzed in this research were historical measurements from weather stations located at six different sites. Future studies may compare the methods for day-ahead forecasts if NWP data is made available.

1.2 Secondary objectives

Some other specific objectives of this research are:

a) Evaluate data processing strategies for intraday solar forecasting;

b) Evaluate clear sky models influence on intraday solar forecasting;

c) Review research on solar forecasting;

1.3 Summary

The thesis content is divided into six chapters. Besides this introductory chapter, the second one brings a literature review; the third one presents the methodology; the fourth and fifth chapters present the results. The sixth chapter brings a general conclusion of the work.

Chapter 2 presents a literature review on solar forecasting research field. An overview on the topic is presented, regarding main types of data acquisition systems that have been adopted, statistical methods that have been developed and error metrics that have been used to assess different forecasts.

Chapter 3 presents the methodology developed to build intraday solar forecasts based on multigene genetic programming (MGGP) modeling and simulation. The statistical methods used as benchmarks and the metrics used to evaluate the proposed methodology are also presented in this chapter.

Chapter 4 brings a first case study comparing MGGP and MLP neural networks for intraday solar irradiance forecasting using large datasets from multiple locations of 3 different countries.

Chapter 5 poses the results of a second case study on developing multivariate intraday solar PV forecasting models based on MLP neural networks and a combination of PV output power measurements and meteorological station measurements acquired from the EMC–UFG facilities.

Chapter 6 closes the text with the author's conclusions and suggestions for future works in the field.

2 Solar Forecasting Research

This chapter brings a comprehensive literature review about the solar forecasting research field and presents the main aspects that have driven methodologies in the field.

2.1 Solar Energy Variability

The potential of solar radiation as a renewable energy source is commonly measured as the variable Irradiance (G) in W/m^2. The variability of Irradiance on a horizontal surface depends on deterministic components and stochastic components. Deterministic components such as day of the year, solar zenith angle, local latitude, time of the day and solar radiation incident on a horizontal plane outside the atmosphere affect solar irradiance incident on the earth's surface. Equations 2.1 and 2.2 describe the influence of time variables in Extraterrestrial Irradiance (G_o) is described in the following Equations (DUFFIE; BECKMAN, 2013):

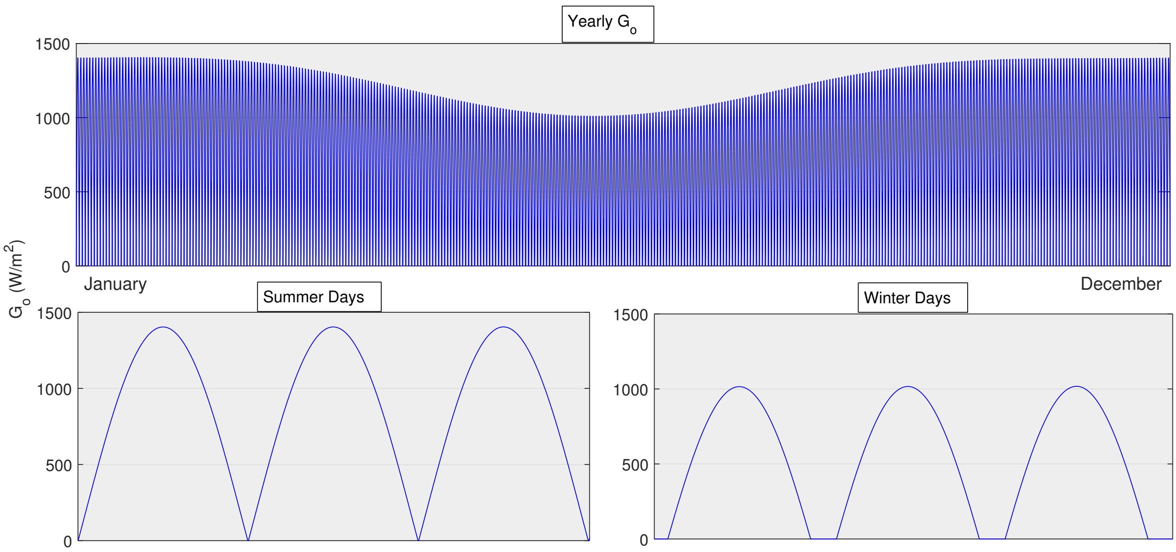
$$G_o = G_{sc}(1 + 0.033 \cos \frac{360n}{365}) \cos \theta_z \tag{2.1}$$

$$G_o = G_{sc}(1 + 0.033 \cos \frac{360n}{365})(\cos \phi \cos \delta \cos \omega + \sin \phi \sin \delta) \tag{2.2}$$

Where G_{sc} is the solar constant, equal to 1367 W/m^2, n is the day of the year (1 for January 1^{st} and 365 for December 31^{st}), θ_z is the solar zenith angle, ϕ is the local latitude, δ is the declination angle, and ω is the hour angle. These equations represent in practice the influence of earth-sun distance and the influence of Earth's rotation movement in the irradiance at a specific horizontal plane at a given location in the globe. The deterministic variability of solar irradiance occurs daily and yearly, as shown in Figure 2.1. During the day, irradiance increases from 0, at sunrise time, to a peak value at solar noon and then decreases to 0 at sunset time. During the year, irradiance levels raise to maximum theoretical values at the summer solstice and reduce to minimum values at the winter solstice.

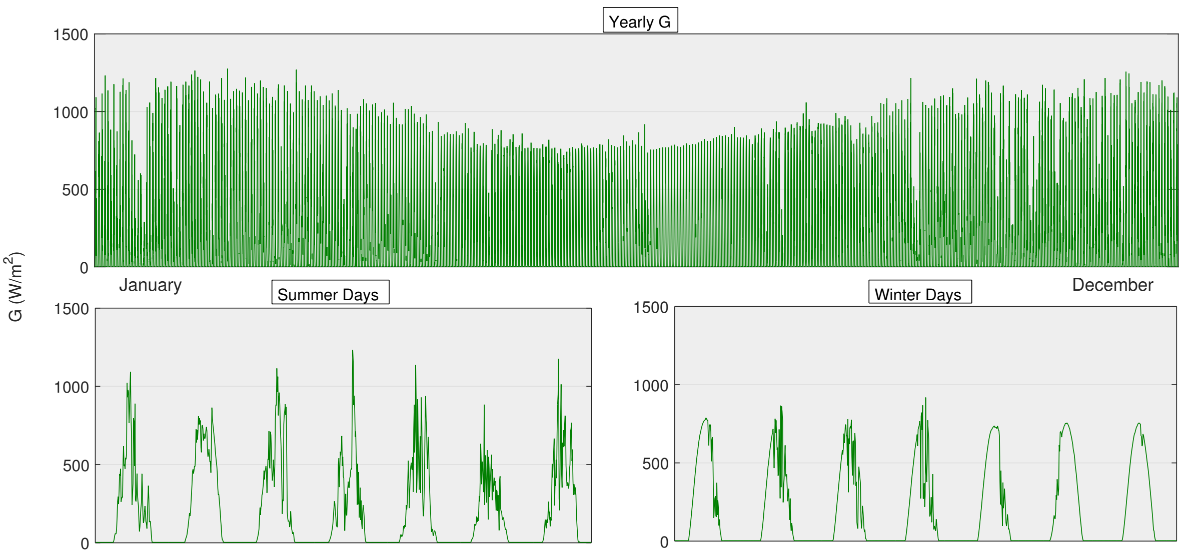
Stochastic variability is the influence of the atmosphere in solar radiation. Scattering and absorption of radiation are atmospheric phenomena that have been estimated by several clear sky models of irradiance at the ground level (DUFFIE; BECKMAN, 2013). Indeed, clouds are the physical phenomena that influence solar irradiance more significantly and generate stochastic time series with significant deviations in solar resources. Solar forecasting studies have tried to model and predict clouds to increase the PV penetration in electrical grids. Figure 2.2 presents global horizontal irradiance (GHI) at ground level

Figure 2.1 – Extraterrestrial irradiance (G_o) in a horizontal plane located in Goiania at an yearly point of view (top) and daily point of view (bottom).



Reference: authors.

Figure 2.2 – Global horizontal irradiance (G) curves from EMC-UFG Weather Station at an yearly point of view (top) and daily point of view (bottom) - year of reference: 2016.



Reference: authors.

from the EMC-UFG weather station. Although the behavior of irradiance on the horizontal ground plane is random, this specific site presents more clear days during the winter and more solar radiation variability during the summer, the classical period of rain days in the Brazilian Midwest.

2.2 Solar Forecasting Reviews

Solar forecasting is a field of research that has been extensively studied in the last years. Firstly, due to its central importance to increase solar energy penetration into electrical grids. Secondly, the complex space-temporal stochastic behavior of solar variability and peculiarities seems not yet promptly predictable for diverse sites.

Solar forecasting is considered the field of research that addresses solar resource predictions, irradiance and PV power, as both quantities are closely related.

Several review articles have been published in the last years in the solar forecasting research field. Review articles intend to compare, present differences and similarities, and classify research articles. Table 2.1 shows leading review articles that are adopted as guidelines for definitions and classifications in this thesis.

Table 2.1 – Some of the main review articles in solar forecasting literature.

Reference	**Article Title**	**Google Scholar Citations**	**Scopus Citations**
(INMAN; PEDRO; COIMBRA, 2013)	Solar forecasting methods for renewable energy integration	708	498
(DIAGNE et al., 2013)	Review of solar irradiance forecasting methods and a proposition for small-scale insular grids	496	347
(VOYANT et al., 2017)	Machine learning methods for solar radiation forecasting: A review	561	395
(YANG et al., 2018)	History and trends in solar irradiance and PV power forecasting: A preliminary assessment and review using text mining	169	134

Source: authors.

Note: Citations numbers were obtained on December 1st, 2020.

In Inman, Pedro and Coimbra (2013), an extensive review of methods used for solar forecasting until 2013 was produced. This review article focused on classifying solar forecasts (irradiance or PV power) using different approaches:

a) Regressive methods, such as autoregresssive (AR), autoregressive moving average (ARMA) and autoregressive integrated moving average (ARIMA), with or without exogenous variables;

b) Artificial Intelligence (AI) techniques, such as Artificial Neural Networks (ANNs) and k-nearest neighbors (k-NN);

c) Remote sensing models, such as satellites;

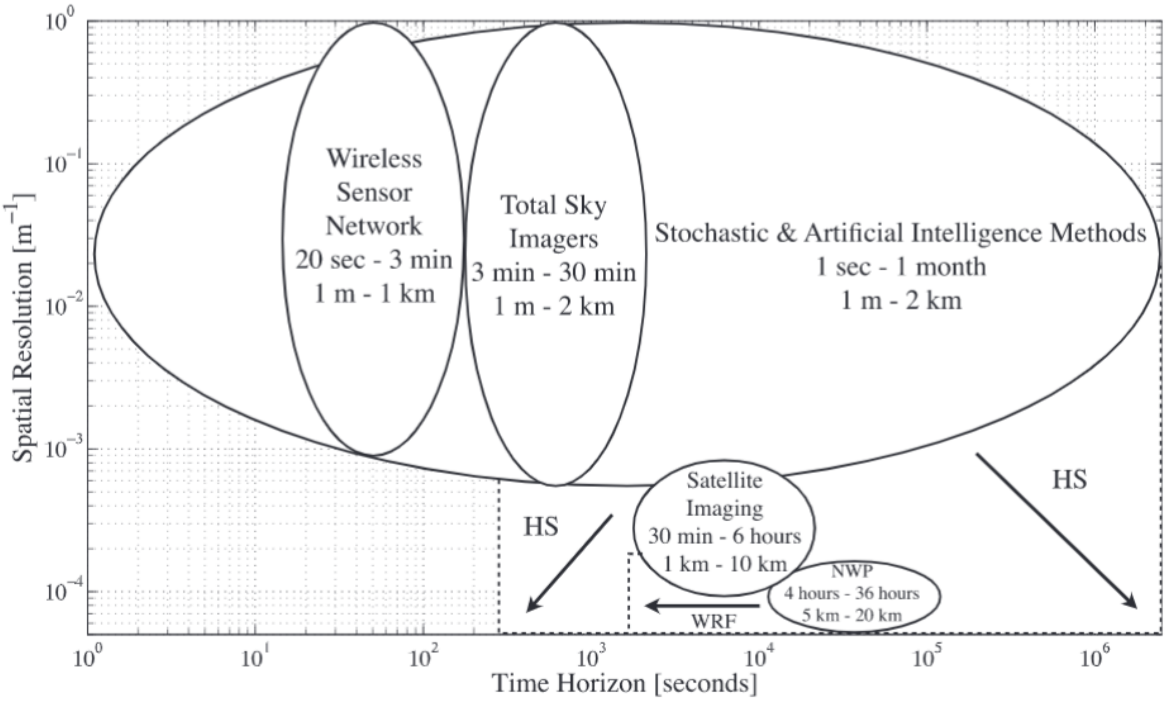
d) Numerical Weather Prediction (NWP);

e) Local sensing, such as sky imagers or pyranometers that directly measures irradiance;

f) Hybrid systems that combine two or more of the previous methods;

The solar forecasting classification by Inman, Pedro and Coimbra (2013) also regards the spatial resolution and forecasting horizon that each technique was able to achieve the best results, as shown in Figure 2.3.

Figure 2.3 – Spatial resolution and forecast horizon domains of different solar forecasting methods by the classification of Inman, Pedro and Coimbra (2013).



Reference: (INMAN; PEDRO; COIMBRA, 2013).

Diagne et al. (2013) published a review on solar forecasting covering papers up to 2013. The survey focused on specific research to predict solar irradiance and the authors classified articles in terms of forecasting horizons as intraday (1h-6h), intra-hour (15-min to 2-h) and day-ahead forecasts (1 day to 3 days) and classified solar predictions in terms of the data acquisition system as:

a) Statistical methods, such as ARIMA techniques and ANNs;

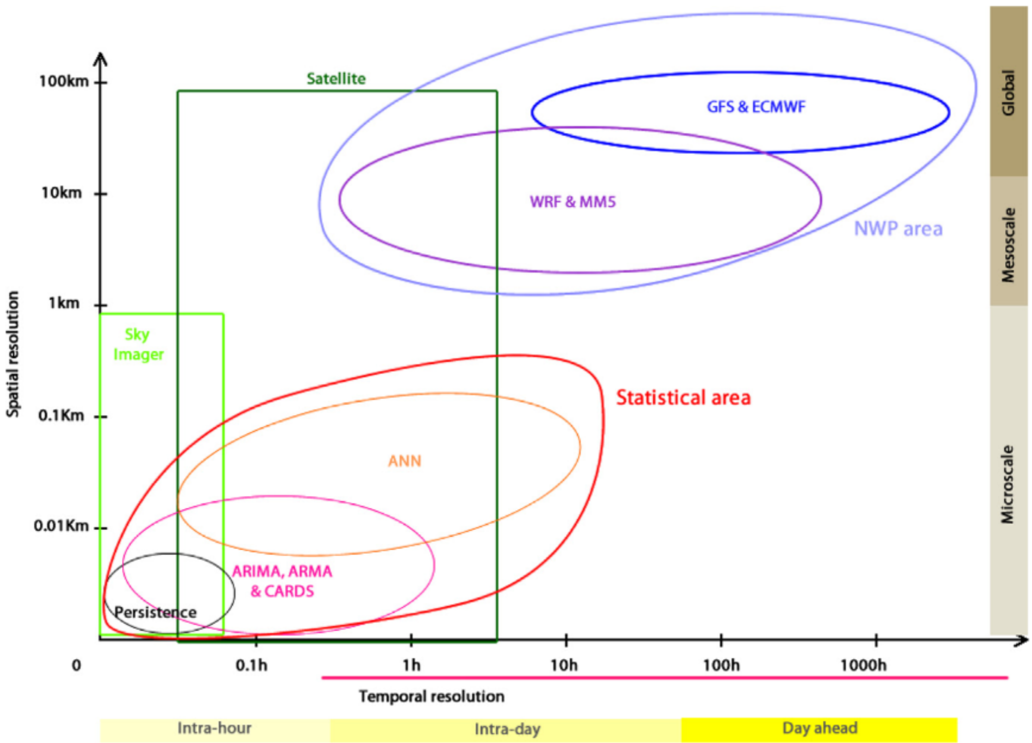
b) Cloud imagery and satellite-based models;

c) Numerical Weather Prediction (NWP) models;

d) Hybrid models;

Diagne et al. (2013) also related the types of solar forecasts with their spatial and temporal areas in which they provide good performances, as shown in Figure 2.4.

Figure 2.4 – Spatial resolution and forecast horizon domains of different solar forecasting methods by the classification of Diagne et al. (2013).



Reference: (DIAGNE et al., 2013).

Yang et al. (2018) presented a review about solar forecasting publications up to 2018 based on a text-mining technique to review a substantial number of research articles, spotting the prominent technical journals in the field and highlighting methods, emerging technologies, and error metrics available. This extensive review considers machine learning as the most adopted statistical method in the area. Voyant et al. (2017) devoted a review exclusively to machine-learning-based solar forecasting.

2.3 Types of data acquisition systems

There is a specific relationship between the type of data acquisition system, forecasting modeling, and the spatial-temporal resolution for each distinct forecasting methodology. This section describes some characteristics of data acquisition systems.

2.3.1 Meteorological stations and power meters

Many proposed solar forecasting methodologies are based on meteorological station data, for solar irradiance forecasting based on pyranometers, or power meter data, for PV power forecasting. While pyranometers present measurement uncertainty ranges of +-

4–7.6%, power meters tend to present lower measurement uncertainties (KLEISSL, 2013). Distributed ground meter networks are going to play an essential role in the future of solar forecasting (YANG et al., 2018).

When forecasts are based exclusively on weather station data or power meter historical measurements, the idea is to develop statistical relations between the recently measured data and the future values of the measured target variable (irradiance or PV power). These are statistical forecasts, according to the definition of Diagne et al. (2013). These kinds of forecasts present its advantages and drawbacks:

a) a) If ground sensors are close to the PV system, they will provide good quality input to forecast models, and the accuracy tends to decay with the distance of the sensor from PV modules, as shown in Figure 2.4.

b) b) Since forecast models rely on last hours' measurements and no future information about the solar resource is available, the accuracy of models is generally limited to one to six hours ahead, as shown in Figure 2.4.

There are two classes of predictions based exclusively on irradiance, weather variables, and PV power measurements: point forecasting and regional forecasting.

Table 2.2 presents a summary of characteristics of some studies on point forecasting in literature based on meteorological stations and power meters. The study of Pedro and Coimbra (2015b) evaluated the performance of statistical methods (k-NN and ANN) for global horizontal irradiance (G) forecasting based on historical data from meteorological stations in the USA. The forecast skill range was 10.2-26% compared to the benchmark persistence method for forecasts horizons with a range of 15–120 min. The models dealt with forecasts of clear sky indexes (normalization of irradiance data) as the first step for irradiance forecasts.

The study of Nobre et al. (2016) evaluated the performance of ARIMA and a hybrid ARIMA-Persistence method to predict irradiance from a meteorological station in Singapore with a forecast horizon of 15–30 min. The study of Rana, Koprinska and Agelidis (2016) evaluated the performance of Support Vector Regression (SVR) and Neural Networks Ensemble (NNens) to forecast power output from a PV plant in Brisbane, Australia, for a horizon of 5-60 min.

There are several studies in the literature addressing statistical point forecasting based on meteorological stations and power meters. Studies differ in how they seek improved forecasts as many factors influence performance, viz:

a) Input or output data time-resolution scheme (intra-hour resolutions contain more information of solar resource and weather variability);

b) Choice of statistical method;

Table 2.2 – Characteristics of data relations of point forecast studies in literature based on meteorological stations for irradiance forecasting (PEDRO; COIMBRA, 2015b; NOBRE et al., 2016) and on PV power meters (RANA; KOPRINSKA; AGELIDIS, 2016). $\Delta\tau$ is the forecast horizon, Δt_{in} is the time resolution of input data, Δt_{out} is the time resolution of output data. MAE and s are error metrics defined in 2.5.

Locations	$\Delta\tau$	Δt_{in}	Δt_{out}	Input Data	Output Data	Results
Merced, Davis, San Diego, Bellingham, Ewa Beach	15	5	15	$k_t^*(t-5,...,t-120)$	$k_t^*(t+\Delta\tau)$	s 12–14.5% (ANN) s 10.7–14.7% (kNN)
	30	5	15	$k_t^*(t-5,...,t-120)$	$k_t^*(t+\Delta\tau)$	s 11.1–16.0% (ANN) s 10.2–16.8% (kNN)
	45	5	15	$k_t^*(t-5,...,t-120)$	$k_t^*(t+\Delta\tau)$	s 10.5–18.1% (ANN) s 11.4–18.0% (kNN)
	60	5	15	$k_t^*(t-5,...,t-120)$	$k_t^*(t+\Delta\tau)$	s 10.7–18.8% (ANN) s 12.0–18.2% (kNN)
	90	5	15	$k_t^*(t-5,...,t-120)$	$k_t^*(t+\Delta\tau)$	s 14.1–21.4% (ANN) s 15.8–23.3% (kNN)
	120	5	15	$k_t^*(t-5,...,t-120)$	$k_t^*(t+\Delta\tau)$	s 15.6–25.7% (ANN) s 17.3–26.0% (kNN)
Brisbane	5	5	5	$P_o(t-5,...,t-840)$	$P_o(t+\Delta\tau)$	MAE 47.72 kW (NNens)
	10	5	5	$P_o(t-5,...,t-840)$	$P_o(t+\Delta\tau)$	MAE 64.48 kW (NNens)
	15	5	5	$P_o(t-5,...,t-840)$	$P_o(t+\Delta\tau)$	MAE 72.28 kW (NNens)
	20	5	5	$P_o(t-5,...,t-840)$	$P_o(t+\Delta\tau)$	MAE 77.70 kW (NNens)
	25	5	5	$P_o(t-5,...,t-840)$	$P_o(t+\Delta\tau)$	MAE 82.05 kW (NNens)
	30	5	5	$P_o(t-5,...,t-840)$	$P_o(t+\Delta\tau)$	MAE 85.84 kW (NNens)
	35	5	5	$P_o(t-5,...,t-840)$	$P_o(t+\Delta\tau)$	MAE 89.25 kW (NNens)
	40	5	5	$P_o(t-5,...,t-840)$	$P_o(t+\Delta\tau)$	MAE 92.67 kW (NNens)
	45	5	5	$P_o(t-5,...,t-840)$	$P_o(t+\Delta\tau)$	MAE 95.04 kW (NNens)
	50	5	5	$P_o(t-5,...,t-840)$	$P_o(t+\Delta\tau)$	MAE 96.67 kW (NNens)
	55	5	5	$P_o(t-5,...,t-840)$	$P_o(t+\Delta\tau)$	MAE 98.59 kW (NNens)
	60	5	5	$P_o(t-5,...,t-840)$	$P_o(t+\Delta\tau)$	MAE 100.23 kW (NNens)
Singapore	15	15	15	$k_t^*(t-15)$	$k_t^*(t+\Delta\tau)$	s 5.8% (Hybrid Pers-ARIMA)
	30	15	15	$k_t^*(t-15)$	$k_t^*(t+\Delta\tau)$	s 5.0% (Hybrid Pers-ARIMA)

References: (PEDRO; COIMBRA, 2015b; NOBRE et al., 2016; RANA; KOPRINSKA; AGELIDIS, 2016).

c) Influence of exogenous input addition (weather variables or deterministic time variables);

d) Choice of clear sky model for analysis;

e) Data pre-processing strategy;

f) Data post-processing strategy;

g) Hybridization strategies;

h) Ensemble strategies;

It is possible to develop a regional forecasting model if a network of distributed pyranometers or power meters over a specific area is available, as Gutierrez-Corea et al. (2016) previously reported. Intraday irradiance forecasting from a meteorological station was improved by adding neighbor stations' information in an MLP-based forecasting model (GUTIERREZ-COREA et al., 2016). The proposed method improved RMSE in 5.89%, 6.97% and 2.55%, for 1-h, 2-h and 3-h ahead forecast, respectively, compared to traditional point forecasting.

2.3.2 Numerical Weather Prediction

Numerical Weather Prediction (NWP) systems are models that directly simulate solar irradiance at multiple levels in the atmosphere and traditionally would not produce predictions of GHI. However, the currently emerging need for solar forecast has required the obtention of GHI from NWP systems (YANG et al., 2018). NWP systems solve fluid motion equations and simulate the atmosphere behavior to achieve a broad spatial and temporal horizon, although this method usually does not provide high spatial and temporal prediction resolutions. One example is the Global Forecast System (GFS), an NWP model available for free access and global coverage. NWP has proved to be a relevant tool for day-ahead solar forecasting, usually providing better results than local sensing for this forecast horizon.

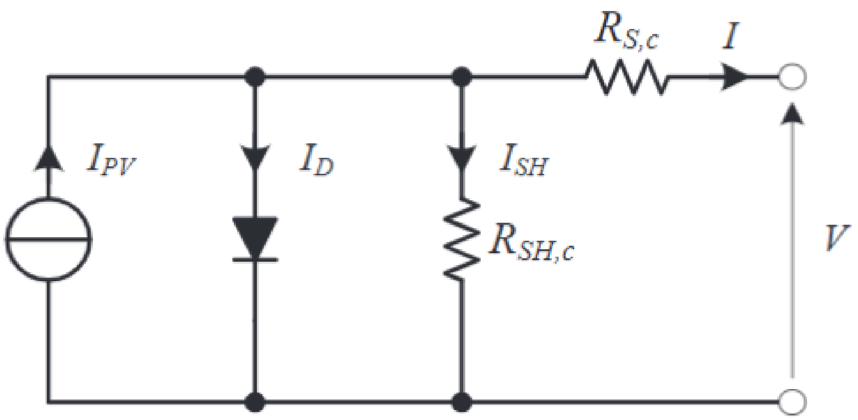
Some other traditional NWP methods have gained prominence in the forecasting research field. Lorenz et al. (2009) used the European Centre for Medium-Range Weather Forecasts (ECMWF) model to develop day-ahead forecasting of distributed PV systems in Germany (Lorenz et al., 2009). Lima et al. (2016) evaluated the accuracy of the Weather Research and Forecasting (WRF) Model for day-ahead forecasting of GHI at multiple sites in the northeast of Brazil (LIMA et al., 2016). According to Yang et al. (2018), the ECMWF model has shown better performance than other models (YANG et al., 2018).

Ogliari et al. (2017) carried out a comparative study between two day-ahead PV-power forecasting methods using NWP-derived data (OGLIARI et al., 2017). The NWP service provides GHI and other weather variables forecasts for each hour of the next day. The first method consisted of forecasts based on deterministic physical models of PV cells applied to weather data derived from the NWP service, as disclosed in Figure 2.5 and Equation 2.4.The second method consisted of forecasts based on a physical hybrid artificial neural network (PHANN) that combined NWP derived data with a theoretical clear sky model to fit values to PV local historical measurements. The second method includes local sensing to improve the first alternative and has reduced the MAE values by 25.13 to 34.55%.

$$I = I_{PV} - I_0(\exp\frac{V + R_{S,C}I}{nV_t} - 1) - \frac{V + R_{S,C}I}{R_{SH,C}} \tag{2.3}$$

Where I is the current in A, I_{PV} is the light-generated current in A, I_0 is the reverse saturation current in A, V is the voltage in V, $R_{S,C}$ is the cell series resistance and $R_{SH,C}$ is the cell shunt resistance.

Figure 2.5 – Deterministic 5 parameters PV equivalent circuit for physical PV power forecasting applied by Ogliari et al. (2017).



Reference: (OGLIARI et al., 2017).

Almeida et al. (2017) used a similar approach employing Random Forest (RF) fitting for PV measurements to improve a parametric methodology based on physical models applied to NWP day-ahead weather forecast (ALMEIDA et al., 2017).

Dolara, Leva and Manzolini (2015) analyzed the use of physical models for PV power estimation based on weather variables. Paiva et al. (2017) also reported physical models that can be similarly applied to PV power estimation (PAIVA et al., 2017).

2.3.3 Satellite data acquisition

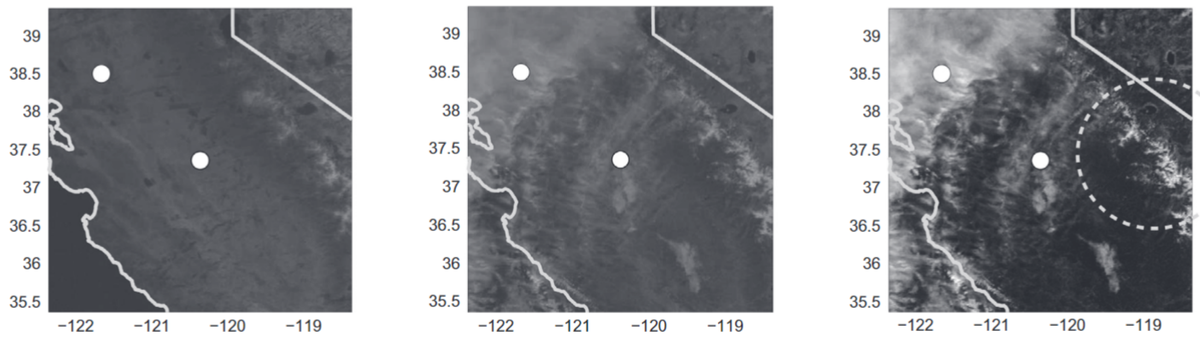
Solar forecasting researchers have applied satellite imagery to develop alternative models for solar forecasting. As a type of remote sensing method, satellite imagery presents the potential to reach large areas as forecast targets. As an example, Pedro, Marquez and Coimbra (2013) developed this technique as expressed in Figure 2.6. The method consists of three steps: 1) Cloud indexing; 2) Image velocimetry; and 3) ANN predictions.

Cloud indexing estimates a pixel intensity by the following equation:

$$E(t) = I_0 \rho \cos\theta(t)^{(1+\alpha)} \tag{2.4}$$

Where $E(t)$ is the pixel intensity, I_0 is the solar constant, ρ is the pixel albedo, θ is the solar zenith angle and α is an empirical parameter determined by trial-and-error. The proposed method was a hybrid satellite imagery-meteorological measurement model and it was evaluated for forecast horizons of 30–120 min.

Figure 2.6 – An example of the cloud indexing procedure. a) Reference ground albedo image obtained by taking the minimum pixel intensities for each pixel location. b) Original image taken on 11/23/2011. c) Cloud indexed image after applying normalization. The two dots identify the forecasting locations. North-most Davis. South-most Merced.



(a) Albedo image (b) Original image (c) Cloud indexed image

Reference: (MARQUEZ; PEDRO; COIMBRA, 2013).

Recent studies evaluated pure satellite-imagery-based intraday forecasting models Miller et al. (2018) concluded that the assumption of cloud invariance is a significant limitation to cloud advection schemes, i.e., the assumption that clouds are not permitted to grow or dissipate during the forecast period yield significant errors and satellite-based forecast techniques require more effective methodologies. Investments in the availability of higher resolution data basis is a promising approach to mitigate the issue.

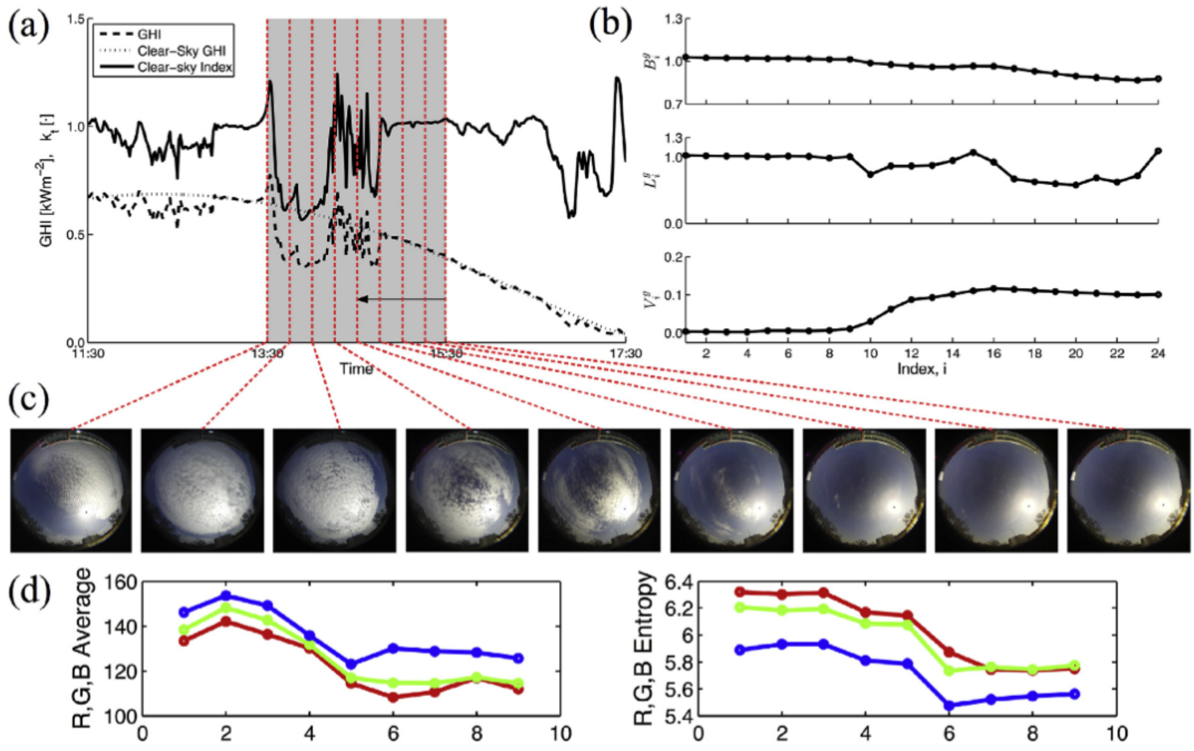
2.3.4 Ground based sky imagers

Ground-based sky imagers are another type of local sensors as pyranometers and power meters adopted for solar forecasting. Instead of directly measuring the solar resource, sky imagers capture cloud cover index and cloud motion behavior to derive solar forecasts.

Pedro et al. (2018) reported results where sky imagers provide potential forecast improvement for forecast horizons of 1-30 min. Pedro et al. (2018) used a security camera that provided 24-bit images compressed in jpeg format, with 8 bits per color channel (Red, Green and Blue) and an overall image resolution of 1563 by 1538 pixels. Figure 2.4 explains how the proposed methodology associates sky images with irradiance measurements to improve solar forecasts.

Pedro et al. (2018) incorporated cloud cover information from sky images, 10 minutes before forecasting time, as inputs to k-nearest neighbors (k-NN) and gradient boosting (GB) algorithms to develop improved intra-hour forecasts. An image processing algorithm computed the average (μ), standard deviation (σ) and entropy (e) for the red, green and blue ratio data sets.

Figure 2.7 – Relations between sky images, cloud detection and irradiance measurements in the methodology of Pedro et al. (2018).



Reference: (PEDRO et al., 2018).

The analysis of sky imager data-addition to GHI forecasts indicated improvements in RMSE values in a range of 0.8–4% using k-NN and a range of 5.2–4.7% using GB for forecast horizons of 5–30 min. The RMSE values for direct normal irradiance (DNI) forecasts improved in a range of 1.3–3.9% using k-NN and 3.8–1.2% using GB for forecast horizons of 5–30 min.

2.4 Statistical methods for solar forecasting

There are several forecasting approaches in the literature; however, the majority rely on statistical techniques to associate data from different acquisition systems to predict solar irradiance or PV power. The most frequent statistical methods applied to solar forecasting literature are described as follows.

2.4.1 Persistence and Smart Persistence

Persistence is the less complex statistical method in the field of solar forecasting. It consists of considering that the future value of the target variable will be equal to its last observed value, as described in equation 2.5:

$$\hat{y}(t + \Delta\tau) = y(t) \tag{2.5}$$

Where $\hat{y}(t + \Delta\tau)$ is the persistence forecast and $y(t)$ is the last observed value of the target variable. When the persistence is applied to the clearness index (k_t) or clear sky index (k_t^*), the method is also called *smart persistence* and equations 2.6 and 2.7. Smart persistence improves a persistence applied to irradiance or PV power in most cases, as indexes are ratios of theoretical solar resource curves. These indexes capture the amount of cloud cover in the last observed time window. A persistence value can also be applied to a clear sky photovoltaic definition, like the k_{PV} defined by Engerer and Mills (2014).

$$\hat{G}(t + \Delta\tau) = k_t(t)G_o(t + \Delta\tau) \tag{2.6}$$

$$\hat{G}(t + \Delta\tau) = k_t^*(t)G_{clr}(t + \Delta\tau) \tag{2.7}$$

Persistence is widely used as a benchmark to evaluate more complex forecast models due to its simplicity, particularly in intraday forecasting. Intelligent persistence demonstrates satisfactory performance to forecast the first future time windows of solar resources, as shown in Figure 2.4.

Not always is the last observed time window the best persistence value. In some cases, the clear sky index from a time window before the last observed one provides more accurate forecast results. An optimized persistence may be used in these cases, as reported by Pedro and Coimbra (2015b).

2.4.2 Autoregressive method

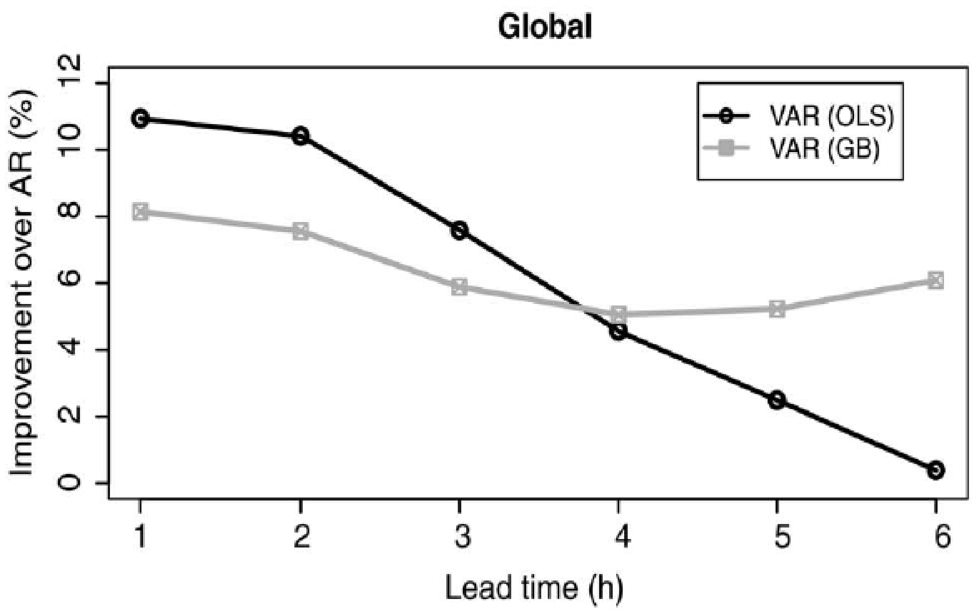
Autoregressive (AR) processes are less complex regression models used in time-series forecasting. The model is defined according to Equation 2.8.

$$\hat{y}(t + \Delta t_{in}) = \alpha + \beta_1 y_{t-\Delta t_{in}} + ... + \beta_l y(t - l\Delta t_{in}) + e_{t+\Delta t_{in}} \tag{2.8}$$

Where $\hat{y}(t + \Delta t_{in})$ is the AR forecast, α is a constant term, β_1 ... β_l are the model's coefficients, l is the order of the AR model and $e_{t+\Delta t_{in}}$ is a white noise with zero mean and constant variance. The regression parameters β_1 ... β_l can be obtained by an Ordinary Least Squares (OLS) methodology, or also by Gradient Boosting (GB). In the last case, a numerical optimization is conducted in order to achieve improved forecasts.

The ARX model is derived from the AR model by adding exogenous variables as inputs. Autoregressive solar forecasting performance becomes more effective when applied for spatio-temporal forecasts, leading to satisfactory improvement over simple AR time-series, as shown by Bessa, Trindade and Miranda (2015). Figure 2.8 presents the improvement of the spatio-temporal forecasting achieved by Bessa, Trindade and Miranda (2015) over AR applied to forecast based on single-site data.

Figure 2.8 – Improvement of Autoregressive (AR) forecasts for 1-6h forecast horizons by application of a spatio-temporal model.



Reference: (Bessa; Trindade; Miranda, 2015).

2.4.3 Autoregressive integrated moving average (ARIMA) models

ARIMA models are the most usual time-series models for solar forecasting (YANG et al., 2018). Equation 2.9 describes the ARIMA(p, d, q) method:

$$(1 - \sum_{i=1}^{p} \phi_i l^i)(1 - l)^d y_t = \delta + (1 + \sum_{i=1}^{q} \theta_i l^i) e_t \tag{2.9}$$

ARIMA models consist of an autoregressive factor of order p, an integrated factor of order d, and a moving average factor of order q. The Box-Jenkings procedure comprises three iterative steps, viz., identification, estimation, and diagnostic checking (BOX et al., 2015), and is the most appropriate way to choose an ARIMA model for a particular forecasting application. When d is null, the ARIMA(p, d, q) model is reduced to an ARMA(p, q) model. If q is null, the ARMA(p, q) model becomes an AR model, and if p is null, the ARMA(p, q) becomes a moving average (MA) model.

Studies that evaluated ARIMA models for intraday solar forecasting have shown similar performances in comparison to smart persistence (NOBRE et al., 2016; REIKARD; HANSEN, 2019).

The main limitation of ARIMA models and time-series, in general, is neglecting to capture the physical behavior of the atmosphere. Therefore, ARMAX models with exogenous weather inputs have improved the pure ARMA and ARIMA models. According to Li, Su and Shu (2014), day-ahead solar forecasts applying ARMAX enhanced ARIMA models improved RMSE and MAE by 26.7% and 28.4% in respect to, respectively.

2.4.4 Artificial neural networks

Artificial neural network (ANN) is the most traditional machine learning (ML) technique in solar forecasting due to its ability to develop non-linear regressions (VOYANT et al., 2017).

Neural networks are connections of neurons grouped in multiple layers. Figure 2.9 presents the neuron structure that relates the output a with inputs $p_1, ..., p_R$. In this case, the inputs are multiplied by the weights $w_{1,1}, ..., w_{1,R}$, and the weighted values are inputs to the sum block with a bias b; finally, a transfer function f is applied to the previous addition to result in the output a, as depicted in Equation 2.10.

$$a = f(\mathbf{Wp} + b) \tag{2.10}$$

Multiple-layer neural networks are as shown in Figure 2.10. In this case, the outputs of each intermediate layer of neurons become inputs to the following layer.

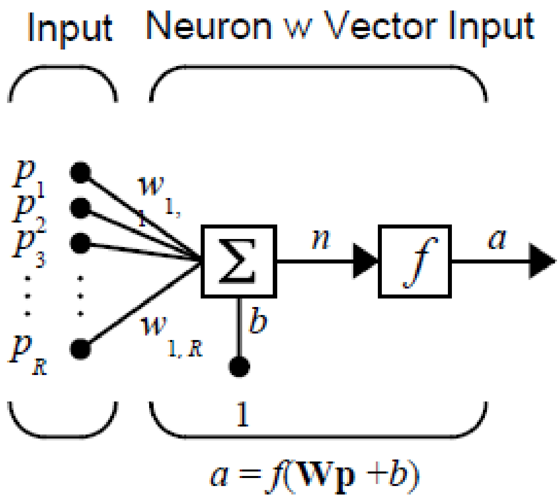
Neural networks are entitled recurrent neural networks (RNN) when the layer's outputs are feedback connected to the inputs.

Elman (1990) developed a neural network called feedforward neural networks (FFNN), where the NN directly connects inputs and outputs without loop iterations. Multilayer perceptron (MLP) and the radial basis function (RBF) neural networks are examples of FFNN.

MLP, as in Figure 2.10, consists of at least three layers of nodes, namely, an input layer, one or more hidden layers with n neurons – 3 hidden layers in the case of Figure 2.10 – and an output layer.

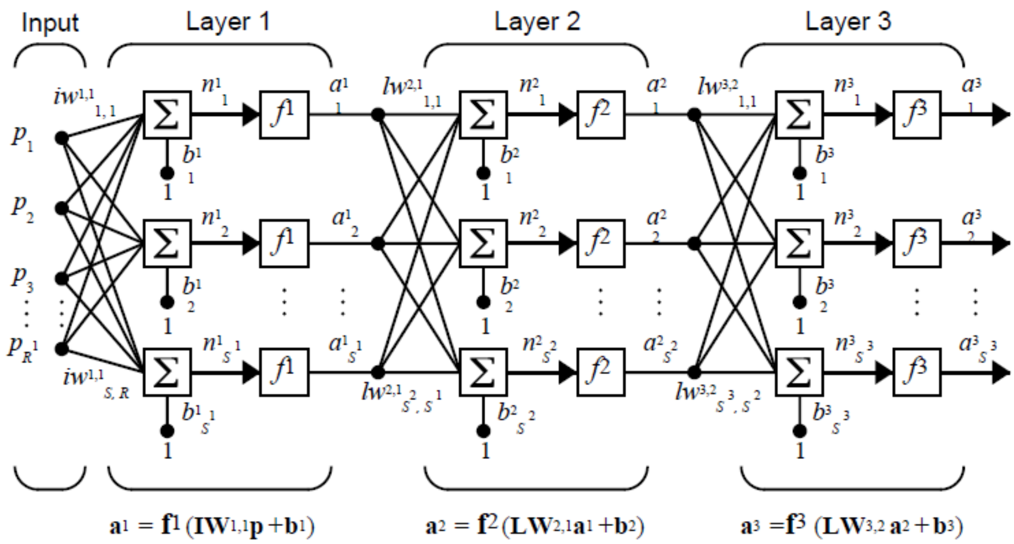
The common characteristic of neural networks and other ML methods is optimizing solutions in three steps, viz., iterative-algorithm training over a specific dataset, testing, and finally validating solutions relying on error metrics. For neural networks, training consists of adjusting the weights and the bias to minimize the output errors. There are

Figure 2.9 – Structure of a single neuron with inputs $p_1, ..., p_R$, weights $w_{1,1}, ..., w_{1,R}$, a bias b and an output a.



Reference: (GUIDE, 2002).

Figure 2.10 – Structure of a multiple-layer neural network.



Reference: (GUIDE, 2002).

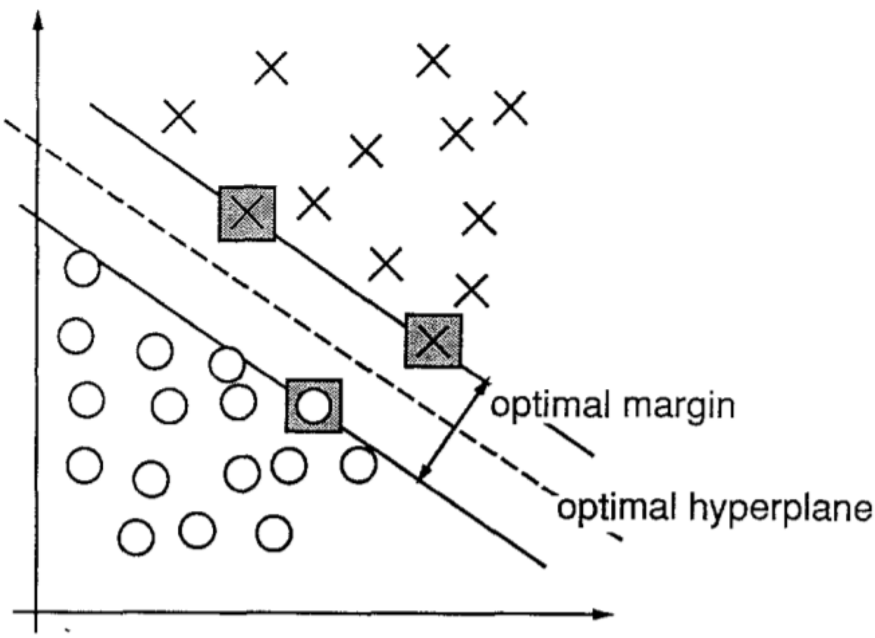
several algorithms for neural network training, such as backpropagation, quasi-Newton, and Levenberg-Marquadt (GUIDE, 2002).

MLP neural networks present the ability to extract information in multivariate solar forecasting models successfully. Numerous implementations combine solar radiation or PV power with different categories of exogenous inputs to improve forecast models.

2.4.5 Support Vector Machines

Support Vector Machines (SVM) are extensively applied methods in solar forecasting and, together with ANN, form the machine learning basis in the field (YANG et al., 2018). Zendehboudi, Baseer and Saidur (2018) presented a review of SVM applications on solar and wind energy (ZENDEHBOUDI; BASEER; SAIDUR, 2018). SVM was originally proposed by Cortes and Vapnik (1995) based on transforming the nonlinear-input area to an area with high-dimensional properties to find a hyperplane via nonlinear mapping. Figure 2.11 presents an example of a separable problem in bi-dimensional space as proposed by Cortes and Vapnik (1995).

Figure 2.11 – Example of a separable problem in a bi-dimensional space by Cortes and Vapnik (1995). Support vectors are marked with grey squares and define the margin of largest separation between two classes.



Reference: (CORTES; VAPNIK, 1995).

If the optimal hyperplane is constructed from a small number of support vectors relative to the training set size, the generalization ability of the SVM will be high, even in an infinite-dimensional space (CORTES; VAPNIK, 1995). Equation 2.11 defines the optimal hyperplane.

$$\mathbf{w}_0 \cdot \mathbf{z} + b_0 = 0 \tag{2.11}$$

Where $\mathbf{w}_0$ are the weights for the optimal hyperplane obtained by a linear combination of support vectors, according to equation 2.12, and the linear decision function in the feature space is obtained from equation 2.13.

$$\mathbf{w}_0 = \sum_{supportvectors} \alpha_i \mathbf{z}_i \tag{2.12}$$

$$I(\mathbf{z}) = sign(\alpha_i \mathbf{z}_i \cdot \mathbf{z} + b_0) \tag{2.13}$$

Where α_i are the adjusted weights of the ith support vector $\mathbf{z}_i$, $\mathbf{z}_i \cdot \mathbf{z}$ is the dot-product between support vectors $\mathbf{z}_i$ and vector $\mathbf{z}$ and b_0 is a bias term.

2.4.6 k-Nearest Neighbors

The k-NN method is one of the most direct and robust available methods of pattern recognition and machine learning and consists of recognizing the instances in the past that represent the current conditions of a target variable as close as possible (YAKOWITZ, 1987).

The most effective forecast model in k-NN methodology relies on comparing the current vector with the time-series past values as defined in equations 2.14 and 2.15 (PEDRO; COIMBRA, 2012):

$$\hat{y}_{t+1} = y_{K+1} \tag{2.14}$$

Where K is such that:

$$\sqrt{\sum_{i=1}(Q_{t,i} - Q_{K,i})^2} \leq \sqrt{\sum_{i=1}(Q_{t,i} - Q_{k,i})^2} \tag{2.15}$$

$k = 1, ..., n;$

$k \neq K$

$$\overrightarrow{Q_t} = (y_t, y_{t-1}, ..., y_{t-\tau}) \tag{2.16}$$

Where $\overrightarrow{Q_t}$ is the current vector of features or patterns and $\overrightarrow{Q_k}$ is the features space.

2.4.7 Random Forest

The Random Forest (RF) belongs to the regression tree (RT) family and is an improved bagging regression tree with promising results for intraday solar forecasting, being more accurate than MLP Neural Networks at some locations (BENALI et al., 2019).

Breiman (1996) proposed the RF method by randomly growing numerous trees in subsets of the predictors.

Equation 2.17 describes the bootstrap aggregating procedure, also known as bagging.

$$\phi_B(x) = av_B \phi(x, \pounds^B) \tag{2.17}$$

Where $\pounds$ is a learning set, x is the input vector, ϕ is a single predictor and ϕ_B is an aggregated predictor.

2.4.8 Gradient Boosted Regression Trees

A promising machine learning method found in solar forecasting literature is the Gradient Boosted Regression Trees (GBRT) that has yielded satisfactory accuracies for intraday PV power forecasting at multiple sites Persson et al. (2017).

The GBRT algorithm consists of an iterative algorithm that adds new regression trees to a fitness function. A fitness function performs a regression to restrict the tree to function residuals in each iteration, as defined in Equation 2.18.

$$F_m(x_{i,t}) = F_{m-1}(x_{i,t}) + h_m(x_{i,t}) \tag{2.18}$$

Where $h_m(x_{i,t})$ is the newly added regression tree at iteration m, equal to the current residuals, which is the negative gradients of the squared error loss function as in Equation 2.19.

$$-\frac{\delta \frac{1}{2}(Y_{i,t+k} - F_{m-1}(x_{i,t}))^2}{\delta F_{m-1}(x_{i,t})} = Y_{i,t+k} - F_{m-1}(x_{i,t}) \tag{2.19}$$

Finally, each regression tree's contribution is weighted by the scaling factor v to avoid overfitting, as shown in Equation 2.20.

$$F_m(x) = F_{m-1}(x) + v h_m(x) \tag{2.20}$$

Where $Y_{i,t+k}$ is the vector of target values in the training set, $F_{m-1}(x_{i,t})$ is the regression model at iteration $m-1$ and $F_m(x_{i,t})$ is the regression model at iteration m.

2.4.9 Comparison of forecasting methods

Voyant et al. (2017) compare forecasting methods from multiple research articles, presenting an overview of machine learning methods from various locations and research groups, as seen in Table 2.4.9. It is possible to perceive comparing methods for distinct locations is not straightforward as a given method presents different performance depending on the site. Moreover, each article refers to diverse forecast horizons, data time-resolution, error-metric definition, type of data acquisition system, and other distinctions.

Table 2.3 – Table of comparison of single machine learning methods for solar forecasting by Voyant et al. (2017).

Reference	Location	Result
Burrows (1997)	Canada	Regression Tree (RT) > Linear Regression
Mihalakakou, Santamouris e Asimakopoulos (2000)	Greece	ANN = AR
Podestá et al. (2004)	Argentina	Generalized regression is useful
Tso e Yau (2007)	China	RT > ANN > Linear regression
Paoli et al. (2010)	France	ANN > AR > k-NN > Bayesian > Markov
Marquez e Coimbra (2011)	USA	ANN > Persistence
Moreno, Gilabert e Martínez (2011)	Spain	ANN = generalized regression
Ben Taieb et al. (2012)	Benchmark	MIMO-ACFLIN strategy wins
Demirtas et al. (2012)	Turkey	k-NN > ANN
Ferrari et al. (2012)	Italy	SVM > ANN > kNN > Persistence
Mori e Takahashi (2012)	Japan	RT interesting to select variables
Olaiya e Adeyemo (2012)	Nigeria	ANN = RT
Fernández, Gala e Dorronsoro (2014)	Spain	SVM = Persistence
Zamo et al. (2014)	Benchmark	RF > SVM > generalized regression > boosting > bagging > Persistence
Almeida et al. (2017)	Spain	Quantile regression forests coupled with NWP give a good accuracy for PV prediction
Wolff, Lorenz e Kramer (2016)	Germany	SVR > k-NN
Lauret et al. (2015)	Fr. Islands	ANN = Gaussian = SVM > Persistence
Lazzaroni et al. (2015)	Italy	SVR > ANN > AR > k-NN > Persistence
McGovern et al. (2015)	Benchmark	RT > NWP
Pedro and Coimbra (2015a)	USA	k-NN > Persistence

Voyant et al. (2017) reviewed articles about ensemble forecasts combining single predictors from one or more statistical methods, generally computing an average of the predictors.

Table 2.4 – Table of comparison of ensemble machine learning methods for solar forecasting by Voyant et al. (2017) (LR = Linear Regression).

Reference	Location	Result
Mori e Kosemura (2001)	Japan	(RT-ANN) > ANN
Cao e Lin (2008)	China	(ANN-Wavelet) > ANN
Reikard (2009)	USA	(ARMA) > ANN
Gastón et al. (2010)	Spain	(SVM-kNN) > Climatology
Chakraborty et al. (2012)	USA	Bayesian > (SVM-ANN)
Bouzerdoum, Mellit e Massi Pavan (2013)	Italy	(SARIMA-SVM) > SARIMA > SVM
Chu, Pedro and Coimbra (2013)	USA	(GA-ANN) > ANN
Prokop et al. (2013)	Czech Rep.	(ANN-SVM) > SVM > ANN
Aggarwal e Saini (2014)	USA	(ANN-LR) > ANN > LR
Alobaidi et al. (2014)	UAE	(ANN) > ANN
Bilionis, Constantinescu e Anitescu (2014)	USA	(PCA-Gaussian Process) > NWP
Wu et al. (2014)	Singapore	(GA-kmean-ANN) > ANN > ARMA
Wu, Chen e Rahman (2014)	Malaysia	(GA-SVM-ANN-ARIMA) > SVM > ANN > ARIMA
Yang et al. (2014)	Taiwan	(ANN-SVM) > SVM > ANN
Chu et al. (2015)	USA	(GA-ANN) > Persistence
De Felice, Petitta e Ruti (2015)	Italy	SVM > Linear Model
Dong et al. (2015)	USA	(ANN-SVM) > ARMA
Samanta, Srikanth and Yerrapragada (2015)	USA	(SVR) > SVR > SVR-PCA > ARIMA > LR

Reported reviews reveal that ensemble predictions typically overcome single predictions, although it cannot be a general conclusion to rank methods applied to distinct locations and with particularities of each circumstance. Nevertheless, additional analyses are needed to reveal conclusive results and rankings.

2.5 Error metrics for solar forecasting

Error-metric definition and interpretation are critical factors in solar forecasting research, as no model is entirely accurate. The error of a single forecast calculated for a single time window is defined by Equation 2.21.

$$e_{(t+\Delta\tau)} = \hat{y}_{(t+\Delta\tau)} - y_{(t+\Delta\tau)} \tag{2.21}$$

Where $e_{(t+\Delta\tau)}$ is the forecast error, $\hat{y}_{(t+\Delta\tau)}$ is the forecasted value and $y_{(t+\Delta\tau)}$ is the observed one.

The reliability of solar forecasting models relies on the analyses of numerous samples, while solar resource and climate are stochastic variables dependent on the specific season of the year and time of the day. The interpretation of single-time windows as in Equation 2.21 is limited to visualization plots, such as in Figure 2.12 (LEVA et al., 2017). The error e defined from Equation 2.21 is normalized, and a percentage error $e\%$ is obtained based on the predicted PV power in Equation 2.22 , and a percentage of the measured PV power is defined by Equation 2.23.

$$e_{\%,p} = \frac{|e_{(t+\Delta\tau)}|}{\hat{y}_{(t+\Delta\tau)}} \tag{2.22}$$

$$e_{\%,m} = \frac{|e_{(t+\Delta\tau)}|}{y_{(t+\Delta\tau)}} \tag{2.23}$$

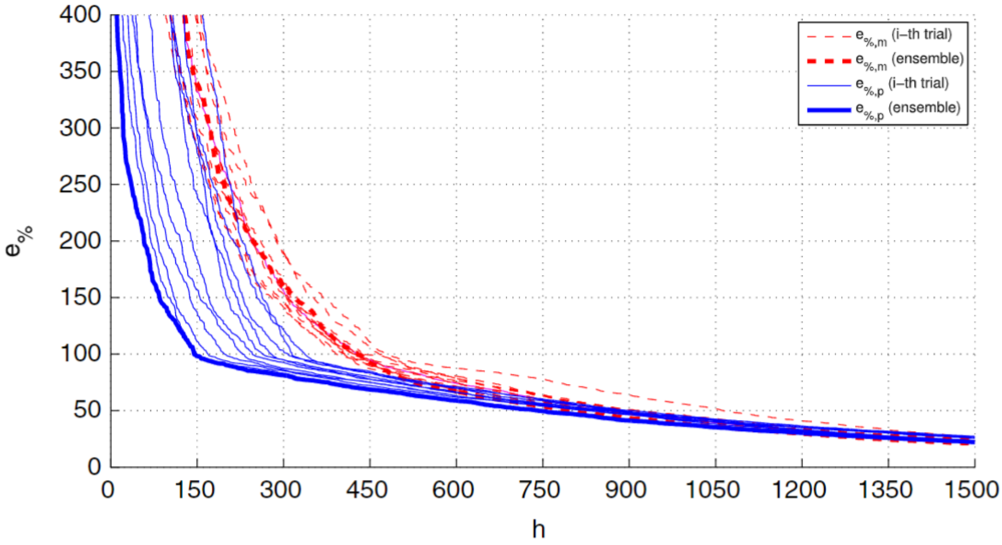
The classic error-metrics for solar forecasting analysis integrate the errors over specific periods of a testing dataset, viz., the mean absolute error (MAE), stated in Equation 2.24, and the root-mean-square error (RMSE), defined by Equation 2.25:

$$MAE = \frac{\sum_{i=1}^{N_{samp}} |(\hat{y}_i - y_i)|}{N_{samp}} \tag{2.24}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N_{samp}} (\hat{y}_i - y_i)^2}{N_{samp}}} \tag{2.25}$$

Where N_{samp} is the number of samples of the specific dataset, y_i is the observed value at sample i and $\hat{y}_i$ is the predicted value at sample i. MAE is the average of absolute errors evaluated for a specific dataset. RMSE has been consistently adopted as error-metric for point solar forecasting (YANG et al., 2018). The advantage of RMSE to

Figure 2.12 – Analysis of day-ahead forecast percentage errors in a sort distribution by Leva et al. (2017).



Reference: (LEVA et al., 2017)

.

evaluate forecast accuracy is that it penalizes higher errors, mainly in RES forecasting, as solar-resource steep ramps are the central issue for application in electrical grids.

A similar definition to the MAE is the mean absolute percentage error (MAPE) that computes the relative gap between predicted and observed values, as established in Equation 2.26:

$$MAPE = \frac{1}{N_{samp}} \sum_{i=1}^{N_{samp}} \frac{|(\hat{y}_i - y_i)|}{y_i} \tag{2.26}$$

The mean bias error (MBE), as in Equation 2.27, gives the average bias between forecast and target variable for the dataset. A negative bias means that the total predicted energy obtained was lower than the total measured energy. Satisfactory MBE should be close to zero.

$$MBE = \frac{\sum_{i=1}^{N_{samp}} (\hat{y}_i - y_i)}{N_{samp}} \tag{2.27}$$

The forecast skill (s) is also a conventional error-metric in solar forecasting, computed as in Equation 2.28. The forecast skill measures the RMSE improvement of a given method relative to the benchmark persistence.

$$s = 1 - \frac{RMSE_{forecast}}{RMSE_{persistence}} \tag{2.28}$$

MAE and RMSE calculations yield absolute values that vary considerably depending on location and climate, making direct comparisons inaccurate. Thus, errors are frequently normalized by the average of observed irradiance or PV power as in nMAE and nRMSE.

$$nMAE = \frac{MAE}{\overline{y_i}} \tag{2.29}$$

$$nRMSE = \frac{RMSE}{\overline{y_i}} \tag{2.30}$$

The Envelope-Weighted MAE (EMAE) and the Objective MAE (OMAE) are also examples of error definitions proposed for PV power forecasting analysis (Leva et al., 2020; Paiva et al., 2019).

$$EMAE = \frac{\sum_{i=1}^{N_{samp}} |\hat{y}_i - y_i|}{\sum_{i=1}^{N_{samp}} max(\hat{y}_i, y_i)} \tag{2.31}$$

$$OMAE = \frac{MAE}{\sum_{i=1}^{N_{samp}} G_{clr}} \tag{2.32}$$

2.6 Summary

This chapter brought a comprehensive literature review on solar forecasting current techniques. While day ahead forecasts usually rely on NWP data, intraday forecasting has chiefly relied on improved statistical methods applied to local sensing. Ranking statistical methods from literature has not been an elementary task, as depicted in section 2.4.9. Recent studies and methods need to be investigated over multiple locations in order to obtain improved forecasts.

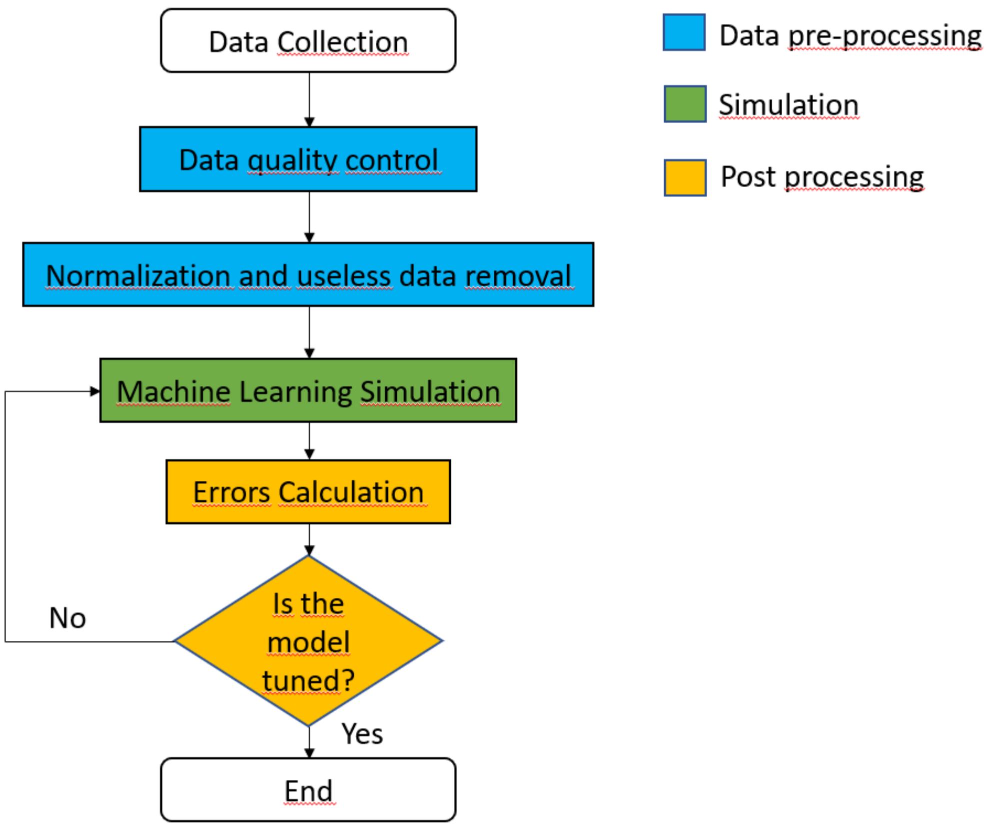
The next chapter reports the methodology of this work and proposes multigene genetic programming (MGGP) as an innovative method for solar prediction. The proposal is validated by comparing simulation results with state-of-the-art methods over multiple location datasets.

3 Intraday Solar Irradiance Forecasting using Machine Learning Algorithms

This chapter describes the proposed methodology to analyze machine learning algorithms in multiple sites.

The flowchart shown in Figure 3.1 summarizes the methodology applied in this research using multiple traditional forecasting methods reported in the literature to obtain improved results. The methodology is composed of three steps: pre-preprocessing to prepare data to the machine learning algorithms; subsequently, multiple machine learning algorithms are simulated to provide enough results to be generalized; finally, postprocessing analyzes error metrics for parameter tuning of machine learning algorithms or to compare methods for solar forecasting at a given horizon and location.

Figure 3.1 – Flowchart of the intraday solar irradiance forecasting methodology.



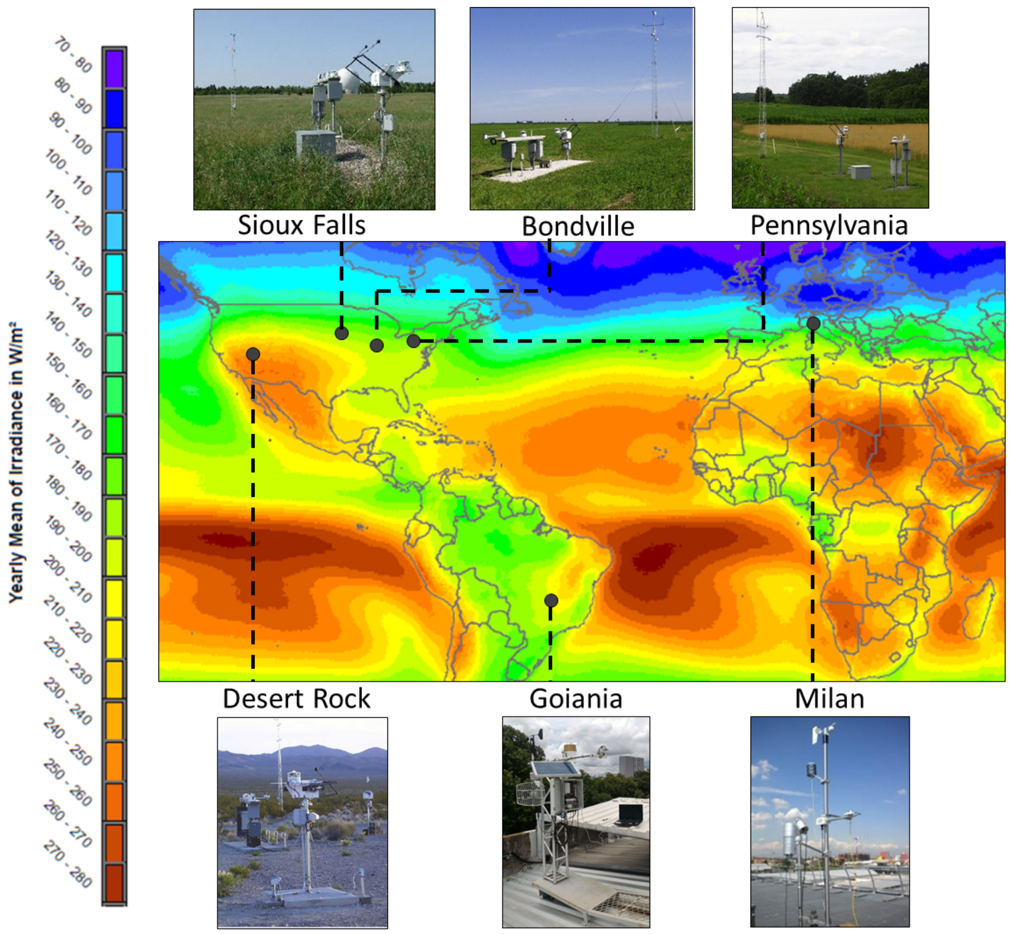
Reference: authors

.

3.1 Data Collection

In this research, databases from 6 meteorological stations were obtained and used to apply the methodology from Figure 3.1. The meteorological stations are located in three different countries: Brazil, Italy and USA, as shows Figure 3.2 (PAIVA et al., 2020).

Figure 3.2 – Locations of the meteorological stations under analysis presented in a world map.



Reference: (PAIVA et al., 2020)

.

3.1.1 Goiania, Brazil

The first database was obtained from the Federal University of Goias (UFG) meteorological station. The station is based at the rooftop of the Electrical, Mechanic and Computer Engineering School (EMC) in the city of Goiania, Brazil. The configurations of the equipments are described in Table 3.1. The coordinates of this station are latitude -16.67° (Southern Hemisphere) and longitude -9.24° (West). The station is located 749 m above sea level (SOLAR. . . , a). The database is a total of 36 months of measurements of weather variables in samples of 1-min resolution from August 2015 to July 2018. The weather variables are: global horizontal irradiance in W/m^2, ambient temperature in °C, atmospheric pressure in mBar, relative humidity in % and wind speed in m/s.

Table 3.1 – Description of the equipment used at the Federal University of Goias (UFG) weather station, the parameters they measure and their accuracy and range of operation.

Equipment	**Parameter Measured**	**Information**
Pyranometer Hukseflux LP02 calibrated	Global horizontal irradiance	Second class ISO 9060: in-field uncertainty of ±5%, calibration uncertainty < 1.8%
R. M. Young Wind 03002	Wind speed	Range 0 to 50 m/s and accuracy of ±0.5 m/s
	Wind direction	Accuracy of ±5%
Texas Electronics TB-2012M	Atmospheric pressure	Calibration range 878 to 1080 mBars, Uncertainty of ±1.3 mBar
Texas Electronics TTH-1315	Ambient temperature Relative humidity	Operating ranges −40 °C–+60 °C and 0–100%, accuracies of ±0.3 °C and ±1.5% RH
Texas Electronics TR-525I	Rainfall	Accuracy of ±1%
Datalogger Campbell Scientific CR800X	Automatic data acquisition	

3.1.2 Milan, Italy

The second database was obtained from the Politecnico di Milano (PoliMi) meteorological station (SOLAR..., b). The station is based at Bovisa Campus, Politecnico di Milano, in the city of Milan, Italy. The coordinates of this station are latitude +45.50° (Northern Hemisphere) and longitude 49.24° (East). The station is located 120 m above sea level. The database is a total of 26 months of measurements of weather variables in samples of 1-min resolution from September 2016 to October 2018. The weather variables are: global horizontal irradiance in W/m^2, ambient temperature in °C, horizontal diffuse irradiance in W/m^2, relative humidity in % and wind speed in m/s.

3.1.3 SURFRAD stations, USA

The third database was obtained from the USA National Oceanic and Atmospheric Administration (NOAA) Surface Radiation Network (SURFRAD). Four meteorological stations from SURFRAD network were chosen to compose the complete database for analysis of this thesis. The coordinates of these sites are as follows: Desert Rock, latitude

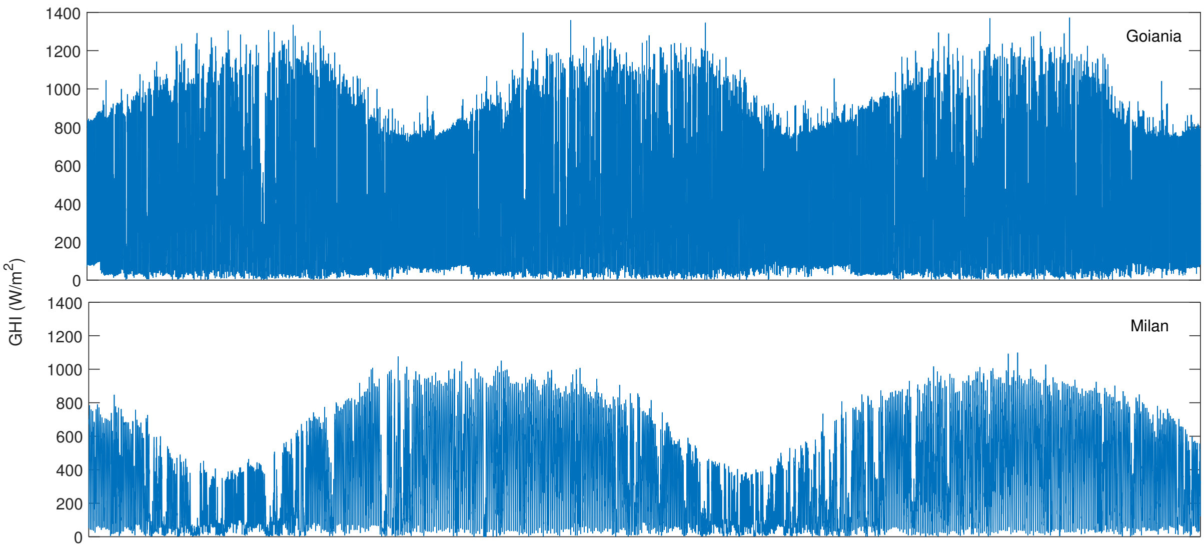
+36.62° (north), longitude -116.02° (west), altitude of 1007 m; Pennsylvania State University (PSU), latitude +40.72° (north), longitude -77.93° (west), altitude of 376 m; Bondville, latitude +40.05° (north), longitude -88.37° (west), altitude of 213 m; and Sioux Falls, latitude +43.73° (north), longitude -96.62° (west), altitude of 473 m. The database is a total of 36 months of measurements of weather variables in samples of 1-min resolution from January 2013 to December 2015. The weather variables are: global horizontal irradiance in W/m^2, ambient temperature in °C, atmospheric pressure in mBar, relative humidity in % and wind speed in m/s.

3.2 Data Pre-processing

3.2.1 Visualization of Irradiance Curves

An overview of the irradiance datasets is presented here in order to relate the obtained results with the characteristics of each dataset and location. Figures 3.3, 3.4 and 3.5 present the irradiance curves of the datasets in 15-min time windows. It is important to notice the diference in peak values of GHI at each season for each location, and also, the 1-year cyclical behavior of GHI increasing from lowest values in winter to highest values in summer. Goiania presents the highest peak values both in summer and winter, because it is located in Southern Hemisphere.

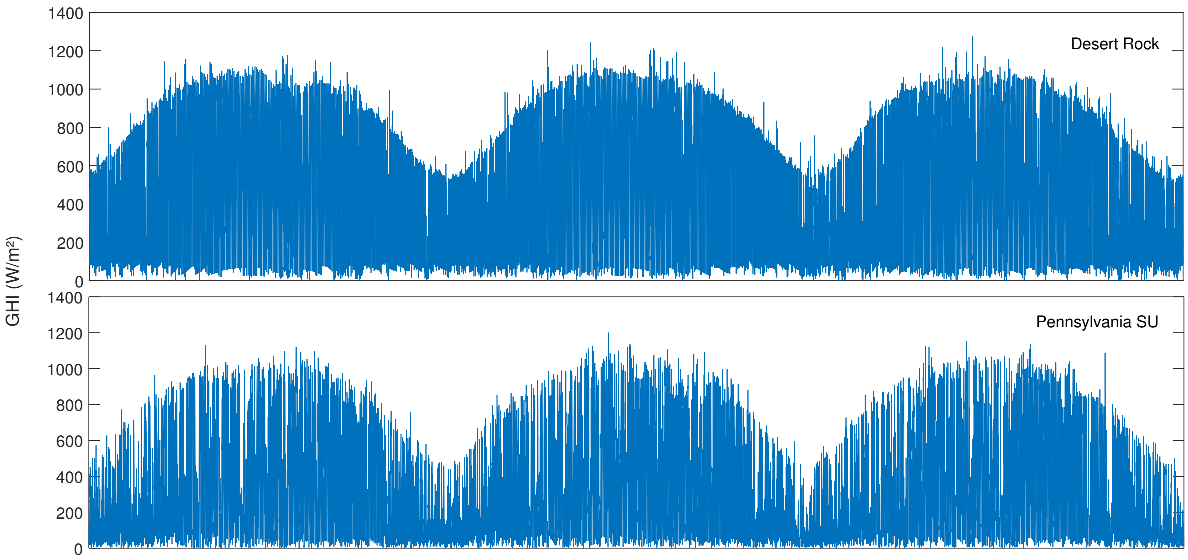
Figure 3.3 – Global Horizontal Irradiance (GHI) curves of Goiania and Milan complete datasets in 15-min time windows.



Reference: authors

.

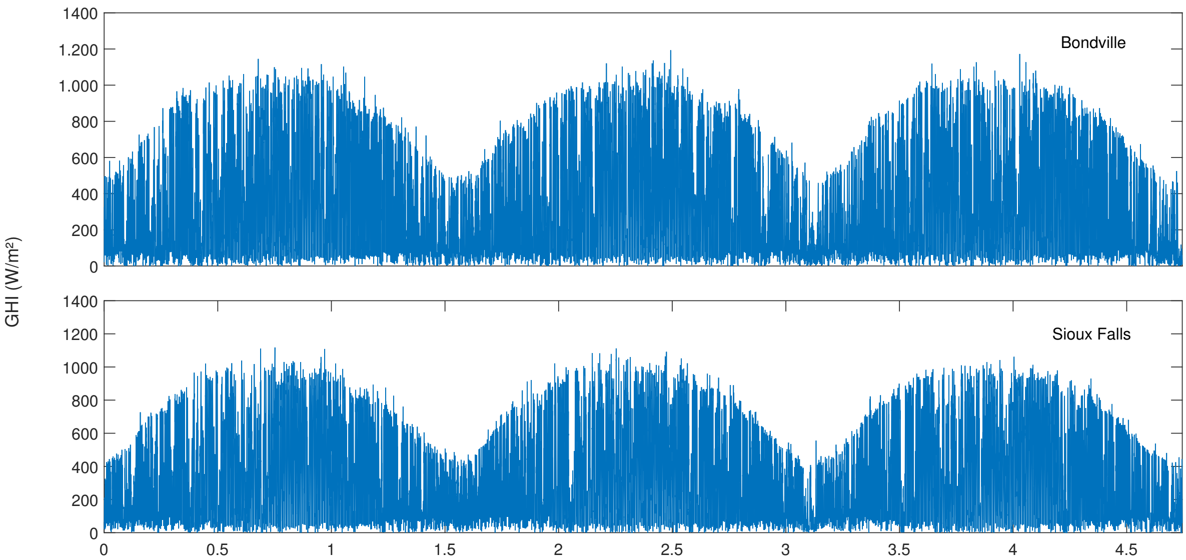
Figure 3.4 – Global Horizontal Irradiance (GHI) curves of Desert Rock and Pennsylvania SU complete datasets in 15-min time windows.



Reference: authors

.

Figure 3.5 – Global Horizontal Irradiance (GHI) curves of Bondville and Sioux Falls complete datasets in 15-min time windows.



Reference: authors

.

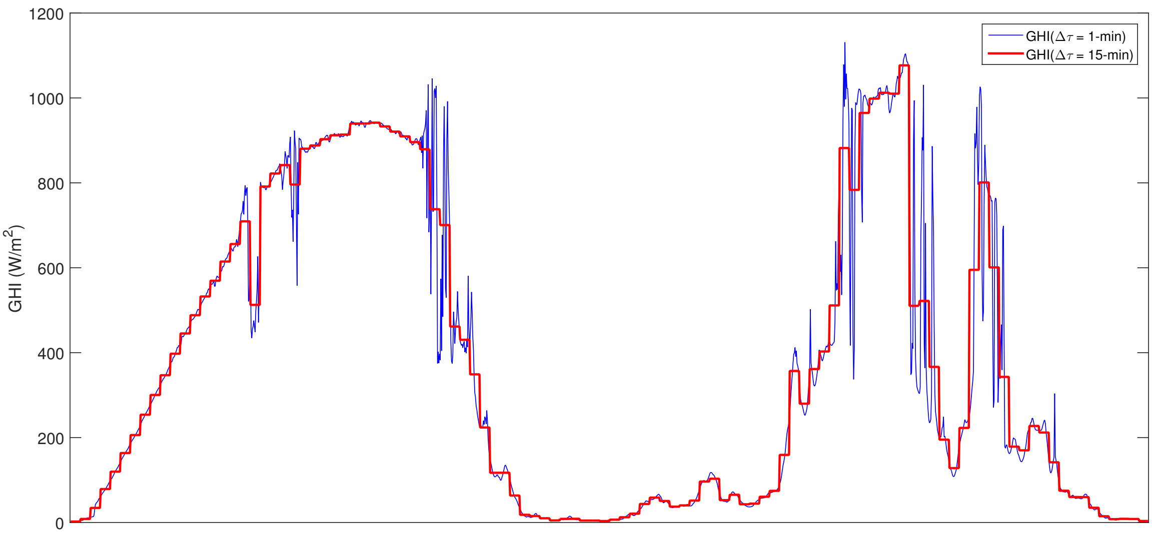
Table 3.2 presents the statistics of GHI per year for each database: total yearly GHI in kWh/m^2, average GHI in 15-min time windows μ_{15} and standard deviation of GHI in 15-min time windows σ_{15}. It is interesting to notice the low variability for each calculated metric from one year to the next one.

Table 3.2 – Statistics of Global Horizontal Irradiance (GHI) for each database: total yearly GHI in kWh/m^2, average GHI in 15-min time windows μ_{15} and standard deviation of GHI in 15-min time windows σ_{15}.

	Year 1			Year 2			Year 3		
	GHI	μ_{15}	σ_{15}	GHI	μ_{15}	σ_{15}	GHI	μ_{15}	σ_{15}
Goiania	1921	465.4	308.9	1903	461.8	309.8	1934	463.1	308.1
Milan	1433	357.2	275.8	1364	344.4	268.5			
Des. Rock	2082	512.9	300.4	2082	512.2	304.2	1954	494.0	299.3
Penn. SU	1317	329.2	266.7	1350	335.9	271.5	1366	339.9	273.9
Bondville	1438	356.5	276.4	1458	364.8	278.6	1438	363.6	274
Sioux Falls	1410	350.9	261.2	1384	354.8	261.4	1437	359.6	263.6

Another important point to notice is that when data is processed in 15-min time windows, the real variability of solar resource decreases, since 1-min measurements are processed into an average 15-min value, as shows Figure 3.6.

Figure 3.6 – Global Horizontal Irradiance (GHI) curves for 2 days in Milan dataset presenting raw 1-min data (blue curve) and processed 15-min time windows data (red curve).



Reference: authors

.

3.2.2 Data Quality Control

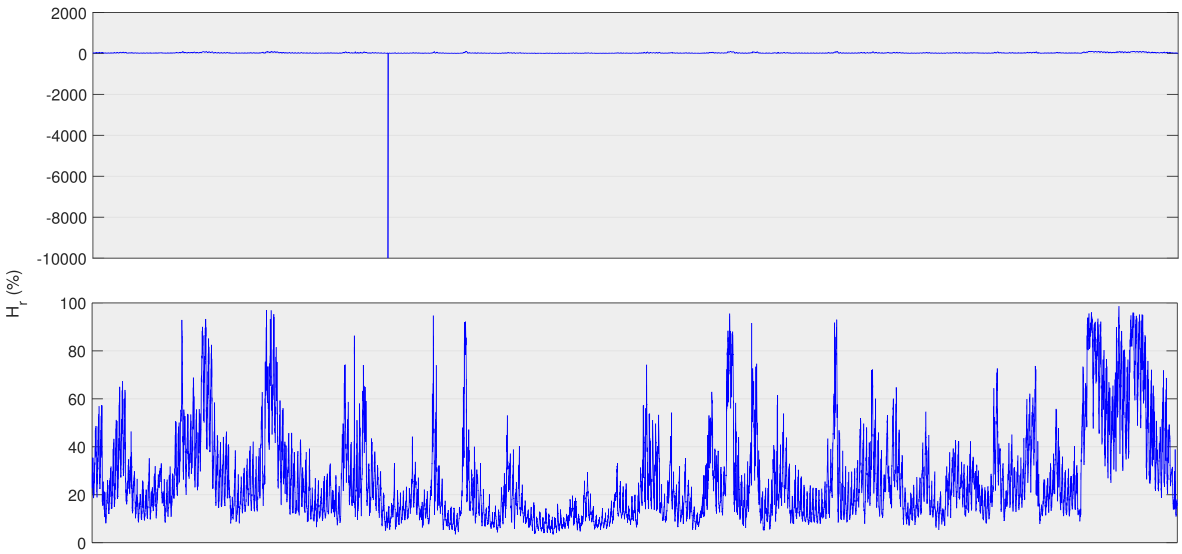
Data quality control is a very important step to build accurate solar forecasts. In this research, first quality control step is applied to remove impossible values in exogenous variables: H_r, T_a, p_a and W_s. The quality control procedure is described by the following code:

Algoritmo 1: Data quality control code

if $(X_t < X_{min})$ *or* $(X_t > X_{max})$ **then**
| $X_t = X_{t-1}$
end

The code consists of applying a persistent value for outliers in weather data. Outliers can significantly reduce the accuracy of machine learning algorithms if not filtered. Figure 3.7 presents an example of a H_r graph after being pre-processed by the data quality control algorithm.

Figure 3.7 – Relative humidity graph of a specific year with an outlier (top) and after being pre-processed by the data quality control algorithm (bottom).



Reference: authors

.

3.2.3 Normalization and useless data removal

[1]Independent of whether forecasts are performed with the use of artificial intelligence methods or classical regressions, the data processing strategy and input–output scheme play a key role in developing improved forecasts. The first data processing strategy considered global horizontal irradiance (G) as a target value, combining past values of irradiance and weather variables in addition to deterministic variables (in order to capture temporal trends in datasets) (de Paiva et al., 2018b; PAIVA et al., 2020).

1 Some texts from this session onwards have been extracted from this research publication version at *Energies* journal (PAIVA et al., 2020).

The proposed approach was refined by adopting a data processing strategy that forecasts normalized indexes in order to remove seasonality in solar data, yielding prompter ML algorithm convergence for irradiance forecasting. Values measured at night and during solar elevations (h) less than 5° were neglected. Normalization of solar data can be performed by the application of Equation 3.1, where k_t^* is the so-called clear sky index, G is the observed global horizontal irradiance (GHI) and G_{clr} is the theoretical clear sky irradiance (PAIVA et al., 2020).

$$k_t^* = \frac{G}{G_{clr}} \tag{3.1}$$

Clear sky irradiance models used in the literature range from simple functions of extraterrestrial irradiance models to complex approaches that take numerous measured atmospheric parameters into account. It was found that Haurwitz clear sky irradiance and Ineichen–Perez models are simple and sufficiently accurate models that were systematically employed to evaluate meteorological data from a wide number of sites in the USA (RENO; HANSEN; STEIN, 2012; PAIVA et al., 2020).

The Haurwitz clear sky irradiance model was developed in 1945 and is given by Equation 3.2, where θ_z is the solar zenith angle (complementary to the solar elevation angle h). The constants 1098 and –0.057 were obtained by the least-squares method in order to fit measured cloudless sky irradiance data from a site in the USA to a theoretical curve based on a zenith angle exponential function. The exponential function is decreased by a factor proportional to $\cos\theta_z$ from sunrise to sunset (PAIVA et al., 2020).

$$G_{clr} = 1098 \cos\theta_z \exp\frac{-0.057}{\cos\theta_z} \tag{3.2}$$

The solar zenith angle is defined as the angle between the zenithal axis and the line to the sun. Thus, this angle varies instantly, according to the rotation movement of the Earth. The cosine of the solar zenith angle is obtained from Equation 3.3, where δ is the declination angle, ϕ is the latitude of the weather station location, and ω is the sun hour angle. A detailed definition and calculation of solar geometry variables is provided in Duffie e Beckman (2013) (PAIVA et al., 2020).

$$\cos\theta_z = \cos\phi\cos\delta\cos\omega + \sin\phi\sin\delta \tag{3.3}$$

Ineichen–Perez clear sky irradiance uses optical air mass ratio (AM), atmospheric turbidity and altitude of location in clear sky irradiance modeling (INEICHEN; PEREZ, 2002). Ineichen–Perez G_{clr} is calculated by Equation 3.4, where G_o is the extraterrestrial irradiance, h is the solar elevation angle, a_1, a_2, f_{h1} and f_{h2} are constant functions of local altitude, T_L is the Linke turbidity factor and AM is the optical air mass ratio. The constants in Equations 3.5 and 3.6 were added empirically by Ineichen and Perez to improve previous clear sky models which were logarithmically dependent on the Linke turbidity factor and limited to specific location and zenith angles. T_L was obtained in this study from a map of monthly averaged values for each site (SODA, 2017). In order to avoid discontinuities in T_L and G_{clr} calculations, a daily fitness procedure was used as presented by Ineichen (2006) and Engerer e Mills (2014) (PAIVA et al., 2020).

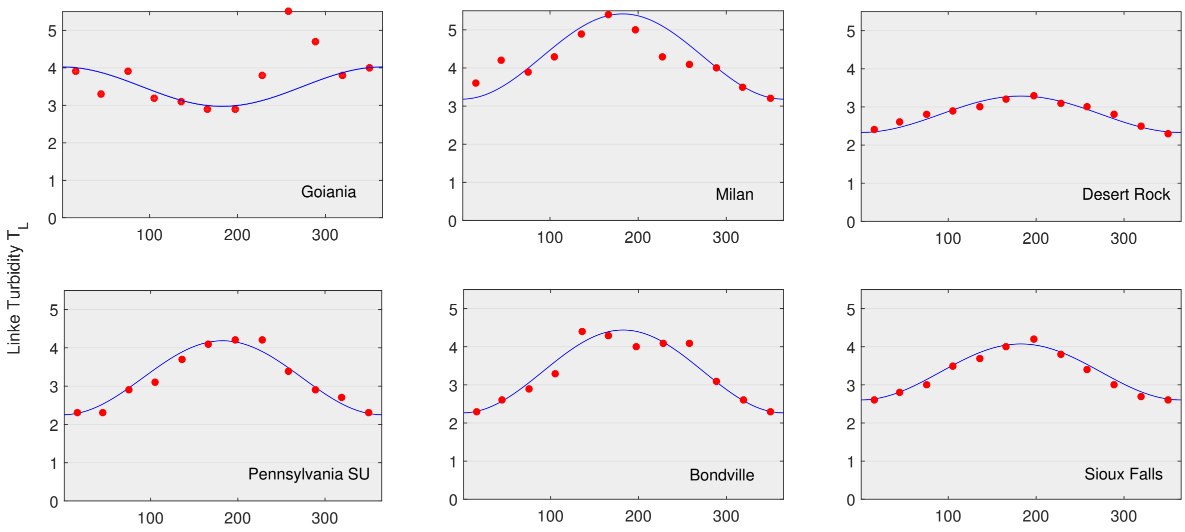
$$G_{clr} = a_1 \cdot G_o \cdot \sin h \cdot \exp[-a_2 \cdot AM \cdot (f_{h1} + f_{h2} - (T_L - 1))] \tag{3.4}$$

$$a_1 = 5.09 \cdot 10^{-5} \cdot altitude + 0.868 \tag{3.5}$$

$$a_2 = 3.92 \cdot 10^{-5} \cdot altitude + 0.0387 \tag{3.6}$$

Figure 3.8 presents the daily fitness procedure applied to each evaluated site.

Figure 3.8 – Linke turbidity daily fitness (blue lines) and monthly averaged values (red dots) for each location.



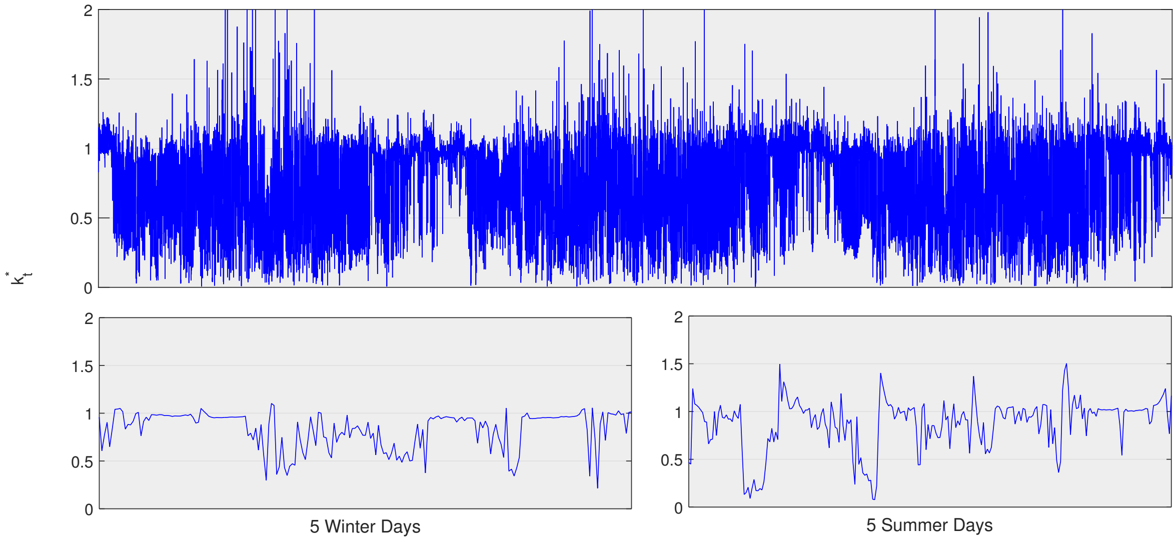
Reference: authors

.

3.2.4 Clear sky index curves and statistics

Figure 3.9 presents the clear sky index graph of the complete dataset of Goiania (15-min time windows), based on Ineichen clear sky model. It is possible to observe how the normalization procedure removes daily and yearly seasonality in solar data. In the case of Goiania, variability is lower in winter, because this region is characterized by dry climate and many clear sky days in winter. Variability is higher in summer, the rainy season in this region.

Figure 3.9 – Ineichen clear sky index curve for the complete dataset of Goiania (top) and for 5 winter and summer days (bottom) (15-min time windows data).



Reference: authors

.

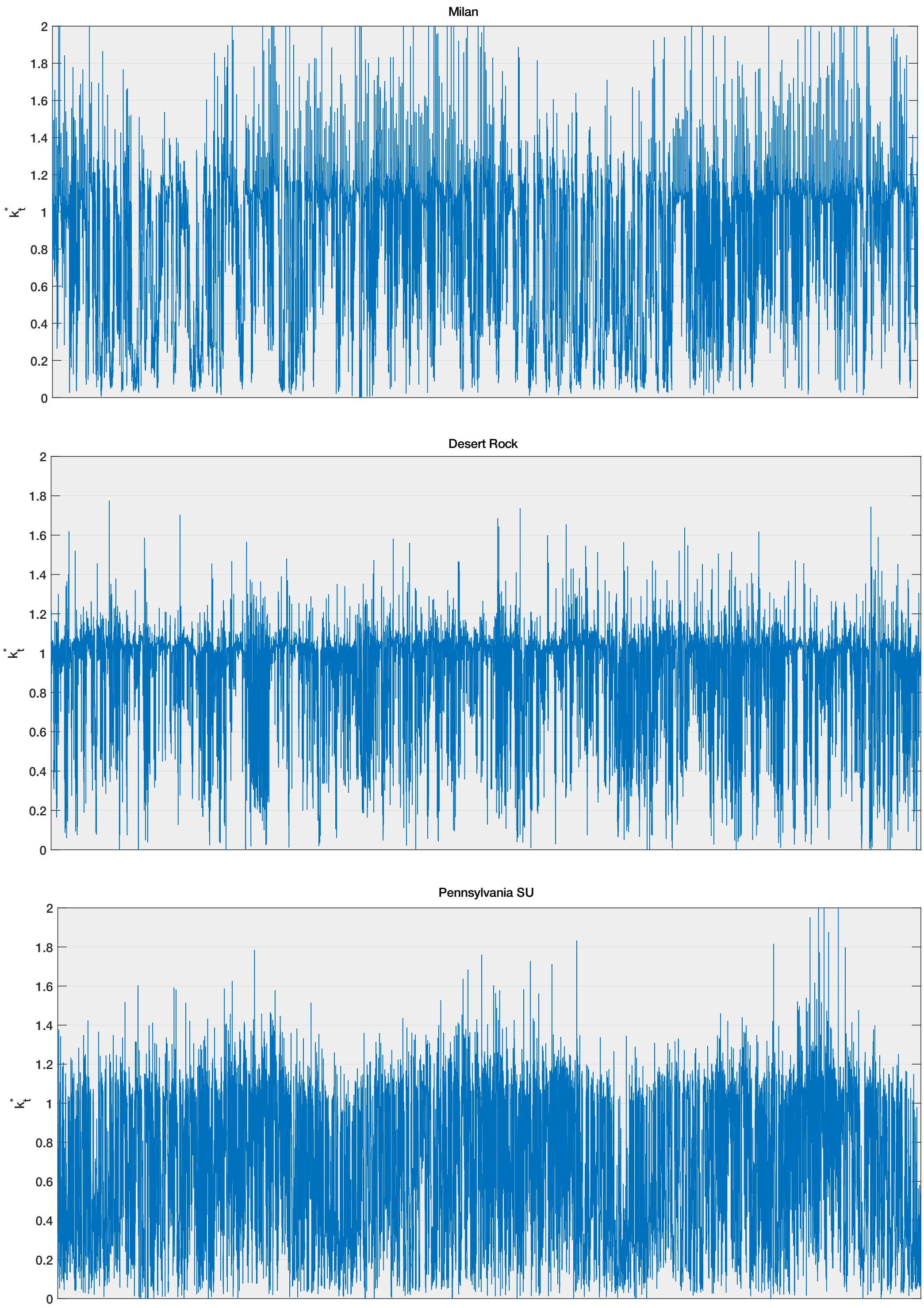
Table 3.3 – Data statistics of training, validation and test datasets for each location: N_{samp} is the number of samples of each dataset, μ is the average Ineichen k_t^* and σ is the standard deviation of k_t^*.

	Train.			**Valid.**			**Test.**		
	N_{samp}	μ	σ	N_{samp}	μ	σ	N_{samp}	μ	σ
Goiania	25,813	0.7379	0.3042	11367	0.7458	0.2983	11,163	0.7423	0.3022
Milan	17,828	0.8544	0.3843	7969	0.8069	0.3897	7944	0.7999	0.4194
Desert Rock	25,959	0.9139	0.2380	10,929	0.9133	0.2451	10,865	0.9025	0.2458
Pennsylvania	25,706	0.6741	0.3534	10,998	0.6260	0.3492	11,177	0.6604	0.3572
Bondville	25,935	0.7246	0.3593	10,818	0.6974	0.3660	11,005	0.7197	0.3478
Sioux Falls	25,839	0.7579	0.3455	10,898	0.7476	0.3594	10,708	0.7638	0.3353

Figures 3.10 and 3.11 presents the clear sky index curves of complete data sets. From top to bottom: Milan, Desert Rock, Pennsylvania SU, Bondville and Sioux Falls, respectively. As seen in Figure 3.9, it is possible to observe how the normalization procedure successfully removes daily and yearly seasonality in solar data. After normalization, both peak and lower values are similar along the years.

Statistics of each site is presented in Table 3.3, achieved by applying the Ineichen clear sky model for 15-min averaged point databases. Results in Table 3.3 show that

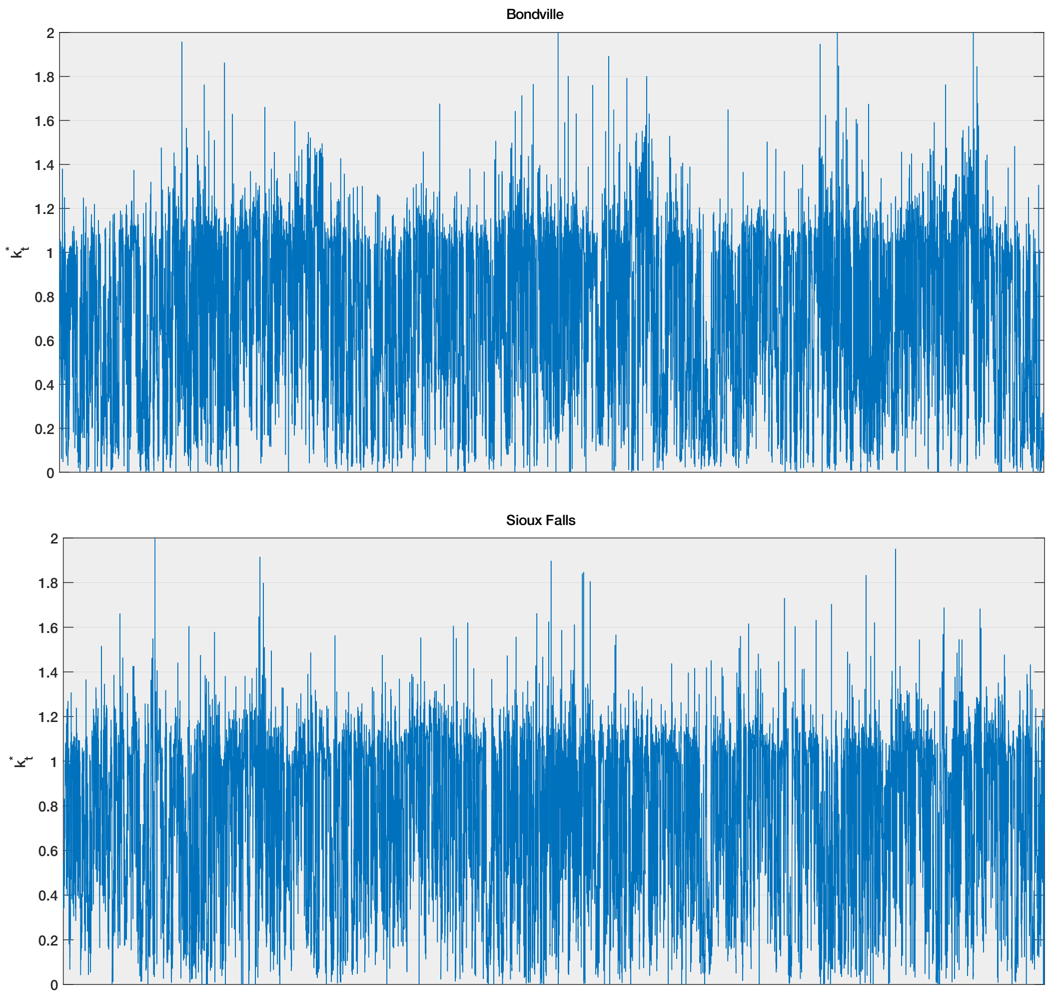
Figure 3.10 – Clear sky index curves of complete data sets. From top to bottom: Milan, Desert Rock and Pennsylvania SU.



Reference: authors

.

Figure 3.11 – Clear sky index curves of complete data sets. From top to bottom: Bondville and Sioux Falls.



Reference: authors

.

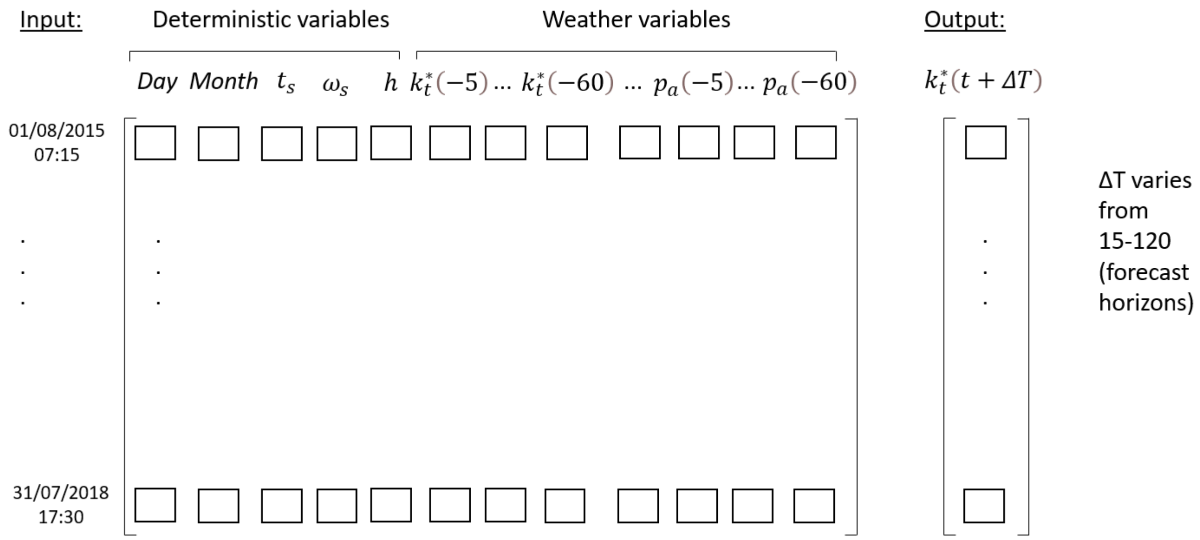
training, validation and testing datasets present similar mean and standard deviations for k_t^*, an important requirement to implement ML forecasting models. Results from Desert Rock present a behavior with more clear sky conditions as opposed to other locations, thus presenting the highest mean k_t^* with lowest standard deviations, while results from Milan present the highest variabilities (σ) (PAIVA et al., 2020).

3.3 Data relations

The ML forecasting methodology is a Multiple Input Single Output (MISO) System. It consists of a "multivariate" data structure of inputs, as defined in (RANA; KOPRINSKA; AGELIDIS, 2016), to forecast k_t^* (single output), using relations among data based on output past values, past values of weather variables and deterministic solar variables. Irradiances are then obtained by multiplying back the normalized index outputs by respective clear sky irradiances (PAIVA et al., 2020).

The MISO system can be visualized in Figure 3.12.

Figure 3.12 – Data relations of the MISO system methodology for intraday solar irradiance forecasting.



Reference: authors

.

A total of 65 inputs are used by ML algorithms to build forecast models for each forecast horizon from 15-min to 120-min ahead. The input variables are:

- k_t^*(-5)...(-60): the 12 past values of k_t^* in time windows of 5 min averages.
- T_a(-5)...(-60): the 12 past values of ambient temperature in °C.
- W_s(-5)...(-60): the 12 past values of wind speed in m/s.
- H_r(-5)...(-60): the 12 past values of relative humidity in %.
- p_a(-5)...(-60): the 12 past values of atmospheric pressure in mBar.
- h is the elevation angle of the forecast time window in radians, varying from around 0.0873 (5°) to 1.5708 (90°).
- t_s is the time difference in respect to sunrise in minutes.
- ω_s is the solar time angle in radians.

- "Day" is the day of forecast interval. The days of the year are counted starting one day after the winter solstice and ending on the winter solstice of the next year. We decided to adopt this definition to follow the solar cycle starting from the day of lowest irradiance levels, since the traditional day counting does not have a direct mathematical relation to the evolution of solar variables throughout the year.

- "Month" is the month of the forecast interval, varying from 1 to 12.

3.4 Multigene genetic programming

Genetic programming (GP) is an artificial intelligence technique which was originally proposed by Koza (1992) in the in the evolutionary computation field; it is considered as an extension of genetic algorithms. GP is inspired by population genetics and biological evolution at the population level (BROWNLEE, 2011) (Algorithm 2). GP has proved to be competitive in time series forecasting in relation to other statistical techniques based on artificial intelligence, such as ANN and the support vector machine (SVM) (LEE; TONG, 2011; GARG; SRIRAM; TAI, 2013; MEHR; KAHYA; OLYAIE, 2013). It has been applied in numerous studies of predictions of natural resources—e.g., hydrology (GHORBANI et al., 2018; MEHR; JABARNEJAD; NOURANI, 2019)—and has also been applied to daily or monthly solar irradiance forecasting in PV power systems (RUSSO et al., 2014; PAN; PANDEY; DAS, 2013; GHIMIRE et al., 2019) (PAIVA et al., 2020).

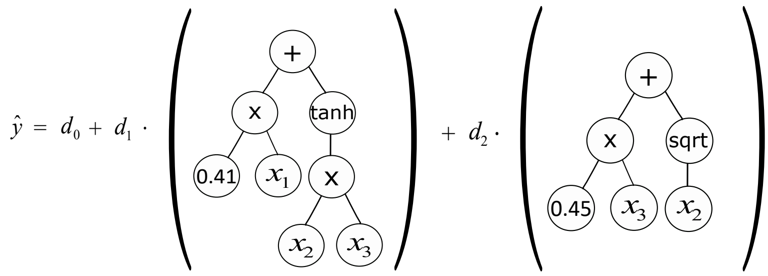
When GP is used to build a mathematical model based on sampled data with the aim of predicting future values, it is named symbolic regression (SR). GP models are typically described as in Equation 3.7, where y is the observed output variable, ŷ is the predicted output, and $x_1 \ldots x_n$ are the observed input variables. In contrast to other soft computing methodologies, such as feed-forward ANNs and SVMs, trained GP models are basic constitutive equations that can be implemented without a specific software environment in any modern programming language (PAIVA et al., 2020).

$$\hat{y} = f(x_1, ..., x_n) \tag{3.7}$$

GP models can be classified into three different categories according to their mathematical model complexity: naive SR, when the model requires only one gene to relate input data with output data; scaled SR, when the model employs one gene associated to a bias term to relate input and output data; and multigene SR, when the GP uses multiple genes and a bias term to relate input and output data (Figure 3.13) (PAIVA et al., 2020).

Figure 3.13 illustrates a population individual and a multigene GP model, usning Equation 3.8, where a bias term d_0 is added to two genes with weights d_1 and d_2 in a tree structure. The terms "plus", "times", "square root" and "tanh" are known as node

Figure 3.13 – Example of a multigene symbolic regression (SR) model presented in a tree structure (PAIVA et al., 2020).



Algoritmo 2: Genetic programming pseudocode (PAIVA et al., 2020)

Input: $Population_{size}$, $Node_{functions}$, $Max_{generations}$, Max_{genes}, Max_{depth}, κ, p_c, p_m, p_r, $elit_{rate}$;
Output: S_{best}
Population $\leftarrow InitializePopulation(Population_{size}, Node_{functions}, Max_{generations}, Max_{genes}, Max_{depth})$;
EvaluatePopulation(Population);
$S_{best} \leftarrow$ GetBestSolution(Population);
while *StopCondition()* **do**
    Children $\leftarrow \emptyset$;
    while *Size(Children)* $< Population_{size}$ **do**
        Operator $\leftarrow$ SelectGeneticOperator(p_c, p_m, p_r)
        if *Operator = CrossoverOperator* **then**
            $Parent_1$, $Parent_2$ $\leftarrow$ SelectParents(Population,$Population_{size}$)
            $Child_1$, $Child_2$ $\leftarrow$ Crossover($Parent_1$, $Parent_2$)
            Children $\leftarrow Child_1$
            Children $\leftarrow Child_2$
        end
        if *Operator = MutationOperator* **then**
            $Parent_1$ $\leftarrow$ SelectParents(Population,$Population_{size}$)
            $Child_1$ $\leftarrow$ Mutate($Parent_1$)
            Children $\leftarrow Child_1$
        end
        if *Operator = ReproductionOperator* **then**
            $Parent_1$ $\leftarrow$ SelectParents(Population,$Population_{size}$)
            $Child_1$ $\leftarrow$ Reproduce($Parent_1$)
            Children $\leftarrow Child_1$
        end
    end
    EvaluatePopulation(Children);
    $S_{best} \leftarrow$ GetBestSolution(Children,S_{best});
    Population $\leftarrow$ Children;
end
return S_{best}

functions. Both weights and nodes are obtained in a GP training procedure (PAIVA et al., 2020).

$$\hat{y} = d_0 + d_1(0.41x_1 + \tanh(x_2x_3)) + d_2(0.45x_3 + \sqrt{x_2}) \tag{3.8}$$

GP evolves a population of candidate solutions (population size) in multiple generations by the application of genetic operators with a tournament selection of best individuals. A crossover operation exchanges genes between individuals to assess possible structural improvements of individuals. Mutation is a fine adjustment operation that changes pieces or entire genes into new, random ones to evaluate a possible structural improvement in terms of fitness. Bias and gene weights of individuals are then optimized in terms of least root mean square errors applied to training data according to Equation 3.9. Applying an elitism strategy with a given elitism rate, a percentage of best fitness solutions is stored over generations. Based on these procedures, GP evaluates thousands of possible regression structures with optimized weights to relate inputs and outputs. Table 3.4 summarizes the parameters adopted in GP, which are considered again in results analysis (PAIVA et al., 2020).

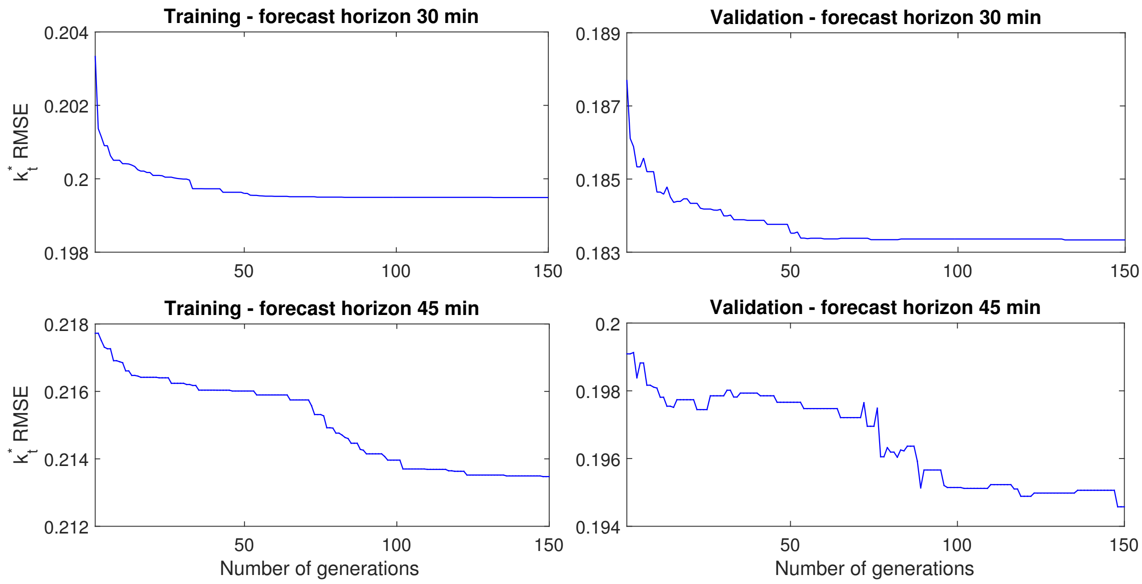
$$s^* = min\sqrt{\frac{\sum_{i=1}^{N_{samp}}(y_i - \hat{y}_i)^2}{N_{samp}}} \tag{3.9}$$

Table 3.4 – Summary of genetic programming (GP) simulation parameters.

Parameter	**Adopted Setting**
Node functions	$+$, $-$, $\cdot$, $/$, x^2, tanh, exp $\sqrt{x}$, exp $-x$, sin, cos
Population size	300
Maximum generations	150
Maximum number of genes	5
Maximum tree depth	4
Tournament size (κ)	6
Lexicographic selection	True
Elitism fraction	0.3
Fitness function	Root mean squared error (RMSE)
Crossover probability (p_c)	0.88
Mutation probability (p_m)	0.12
High-level crossover probability	0.5
Ephemeral random constants range	from -10 to $+10$
ERC probability at creating nodes	0.2

The dynamics of GP solutions are characterized by generalization ability, providing both accurate and robust solutions in training and for other datasets. On the other hand, ANN is highly influenced by overfitting, which is usually controlled by a validation step named early stopping, while GP does not require a validation step during the SR model training stage. Figure 3.14 presents the performance of the best individuals which evolved over generations for GP forecasts. It is possible to observe the robustness of solutions repeating from training to validation datasets. MGGP models were implemented on GPtips 2—an open-source GP platform for Matlab® (SEARSON; LEAHY; WILLIS, 2010) (PAIVA et al., 2020).

Figure 3.14 – Fitness of best GP solution s^* measured by k_t^* RMSE for training and validation datasets (PAIVA et al., 2020).



3.5 Multilayer perceptron (MLP) Artificial Neural Networks

A feed-forward multilayer perceptron neural network (MLP) architecture was applied to this analysis, containing one hidden layer of 10 neurons using the hyperbolic tangent sigmoid transfer function. The neural networks were trained with the Levenberg–Marquadt algorithm including early stopping implemented in Matlab® using the neural networks toolbox. The employed ANN set of attributes was previously validated for intraday solar forecasting (PEDRO; COIMBRA, 2015b) (PAIVA et al., 2020).

3.6 Ensemble Forecasts

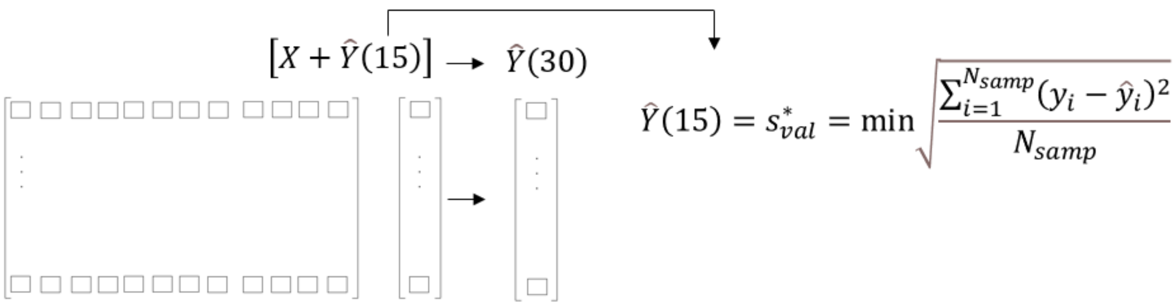
Ensemble forecast models are convenient to build with multiple ML simulations and tend to improve forecast accuracy (LEVA et al., 2017). The ensemble forecast in this research is given by Equation 3.10, where N_{trial} is the number of trials by the given ML method. In this analysis, the internal parameters of GP and ANN do not vary in each trial, and 10 trials were performed to produce each ensemble according to the methodology described in Leva et al. (2017) (PAIVA et al., 2020).

$$\hat{G}_{ens} = \frac{\sum_{i=1}^{N_{trial}} \hat{G}_i}{N_{trial}} \tag{3.10}$$

3.7 Iterative Forecasts

Rana et al. (RANA; KOPRINSKA; AGELIDIS, 2016) evaluated a forecast method where predictions of instant t+1 are iteratively added as inputs to predictions of instant t+2. As a conclusion, the iterative method did not improve forecasts in their study on PV power forecasting using ANN ensemble and SVM; however, the iterative GP method was tested in this work and yielded improvements on forecasting results. Results comparable to (RANA; KOPRINSKA; AGELIDIS, 2016) were obtained, and no significant improvement was achieved by using iteration for ANNs. Therefore, the results reported here were obtained using iterative predictions for MGGP (PAIVA et al., 2020). Figure 3.15 presents data relations in iterative predictions.

Figure 3.15 – Data relations in iterative forecasts.



3.8 Persistence

The smart persistence was also implemented as benchmark technique. Persistence forecasting was computed by Equation 3.11, where $\hat{G}(t + \Delta T)$ is the persistence forecast and ΔT is the forecast horizon, which varies from 15 to 120 min; $k_t^*(t)$ is the present clear sky index; and $G_{clr}(t + \Delta T)$ is the clear sky irradiance at the horizon of the forecast (PAIVA et al., 2020).

$$\hat{G}(t + \Delta T) = k_t^*(t) \cdot G_{clr}(t + \Delta T) \tag{3.11}$$

3.9 Summary

This chapter presented the datasets analyzed, the MGGP algorithm, and the methods to improve forecast models. Data quality control, choice of data structure, ML algorithm development, and tuning need special attention to optimize forecasting models.

The next chapter reports the first-case study results of intraday solar irradiance forecasting evaluated at 6 locations in Brazil, Italy, and the USA.

4 Case study I: Comparison of MGGP and MLP Neural Networks for intraday irradiance forecasting at multiple sites

The first case results are presented in this chapter, addressing intraday solar irradiance forecasting based on data from six meteorological stations situated in Brazil, Italy, and the USA.

4.1 MGGP tuning

Initial simulations were intended to analyze the influence of GP parameters in forecast accuracy and robustness. The analysis of parametric influence is known as the parameter tuning of evolutionary algorithms (EAs), as described in (EIBEN; SMIT, 2011). Parameter tuning is by nature an optimization task comprising multiple variables (parameters). In current analyses of multiple horizon forecasts, each forecast horizon at each location consists of a different problem to be tuned. In order to reduce the number of simulations to assess GP parameters, this study considered prior knowledge from other studies to seek good parameter choices to perform a lower number of simulations. Therefore, parameter assessment was carried out only for one forecast horizon using the dataset from Goiania station. Thus, parameter settings from Goiania were used in forecasting models for other sites (PAIVA et al., 2020).

Lima et al. (de Lima et al., 2010) performed a systematic analysis of GP that indicated the population size, number of generations and tree size as the main parameters which influence fitness, while genetic operators have a lower influence. Increases in the size limit of regression functions tend to improve fitness; however, when the size limit is excessively large, this leads to a bloat (function size growth without fitness improvement) (POLI; LANGDON; MCPHEE, 2008). Bloat can be relieved by using realistic elitism rates (POLI; MCPHEE; VANNESCHI, 2008). In summary, lower tournament sizes and lower elitism rates lead to a higher diversity of solutions (PAIVA et al., 2020).

According to the literature review and some former analyses of irradiance forecasting, the maximum number of genes was set at 5, the tree depth at 4, the number of generations at 150 and population size at 300. These parameters presented a good trade-off between complexity and fitness improvement. Figure 4.1 presents the improvement of solution fitness in the validation dataset from Goiania station versus the increase in complexity (increasing the maximum number of genes) (PAIVA et al., 2020).

Genetic operators were analyzed by multiple simulations for a forecast horizon

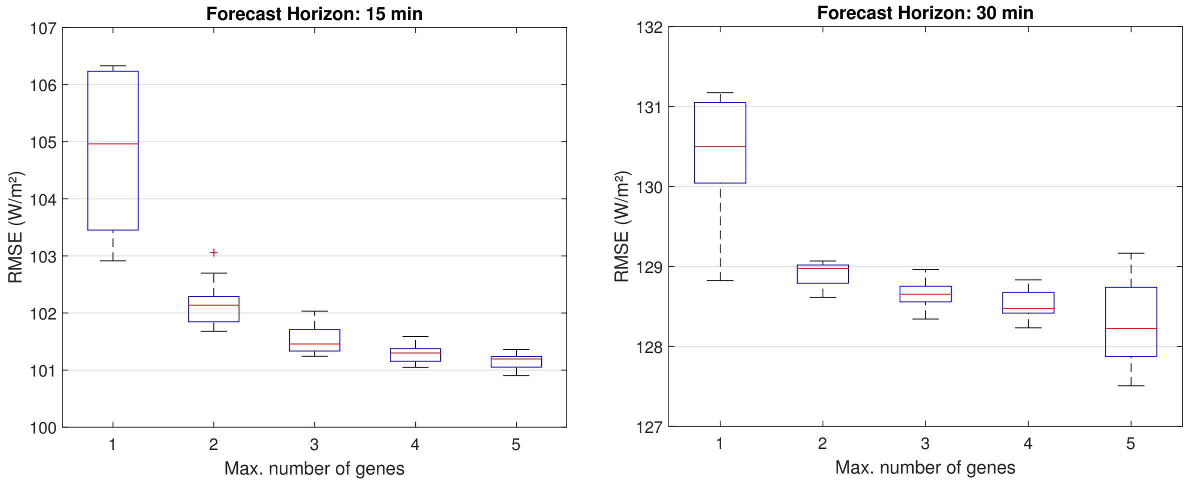


Figure 4.1 – Influence of maximum number of genes on the fitness of best solutions, evaluated for the Goiania validation dataset (PAIVA et al., 2020).

of 90 min ahead, as this is a demanding time window for prediction and consequently presents high variability in the different algorithm simulations. The results for accuracy and robustness are given in Figure 4.2. The number of generations was lowered to 50 during tests in order to obtain a higher variability of results. It is possible to conclude that the best accuracy and robustness (standard deviation for multiple simulations) were those accomplished using higher mutation rates, lower tournament sizes and lower elitism rates. Therefore, we selected the setting with lowest RMSE: $\kappa = 6$, $p_m = 0.12$ and elit $= 0.30$ (PAIVA et al., 2020).

RMSE (W/m²)

κ	pm	elit		
		20	30	40
4	4	154.08	153.76	154.34
6	4	153.38	153.85	154.07
8	4	153.85	154.51	154.60
4	8	153.43	153.85	153.83
6	8	153.68	153.97	154.57
8	8	153.71	154.00	153.94
4	12	153.02	153.55	153.62
6	12	152.99	**152.96**	153.69
8	12	153.51	153.23	153.68

σ_{RMSE}(W/m²)

κ	pm	elit		
		20	30	40
4	4	2.0915	1.4118	1.8041
6	4	1.3997	1.4279	1.7804
8	4	1.8971	2.2814	1.9109
4	8	1.6731	1.7005	1.3713
6	8	1.5721	1.8785	1.6551
8	8	1.4685	1.6325	1.5609
4	12	1.0850	1.4511	1.4577
6	12	1.5722	1.2615	1.6763
8	12	1.4465	**1.0450**	1.8219

Figure 4.2 – Influence of tournament size (κ), mutation rate (p_m) and elitism rate (elit) on the accuracy and robustness (RMSE standard deviation) of the validation dataset from the Goiania site (PAIVA et al., 2020).

4.2 Assessment of exogenous input variables

ANN and GP were executed for all formerly defined locations and forecast horizons both considering and neglecting weather variables H_r, T_a, W_s and p_a. The error improvement index, $Improv_{error}$, was defined in Equation 4.1 in order to assess the improvement yielded by the addition of weather variables at a given error metric, where $error_{univ}$ is the forecast error obtained based on past values of k_t^* with the sole addition of deterministic variables, and $error_{multiv}$ is the forecast error obtained by including weather variables. It is worth highlighting that deterministic variables are able to improve forecasts based merely on past values of k_t^* (PAIVA et al., 2020).

$$Improv_{error} = \frac{(error_{univ} - error_{multiv})}{error_{univ}} \cdot 100\% \tag{4.1}$$

Improvements were calculated both in terms of MAE and RMSE, as described in Figure 4.3. The graphs represent typical behaviors, where weather variables generally improve forecastability for all locations by up to 5.68% in terms of MAE and 3.41% in terms of RMSE; in some locations, negative improvements were obtained for shorter forecast horizons from 15 to 60 min. Mostly, the addition of weather variables tends to improve forecastability for all locations; thus, the results obtained by the multivariate forecasts are reported (PAIVA et al., 2020).

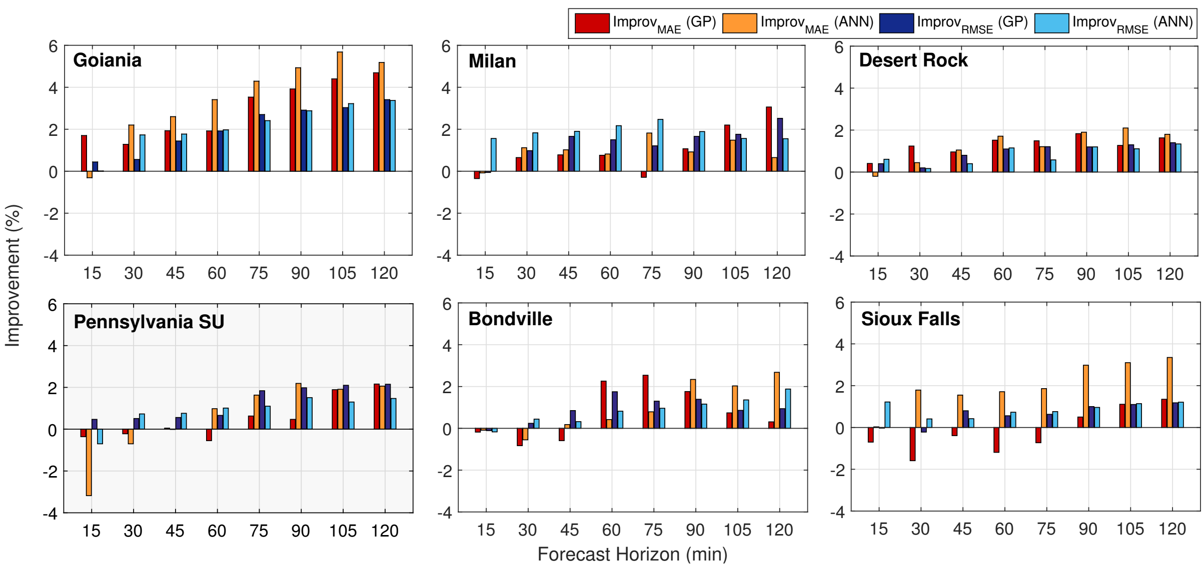


Figure 4.3 – Improvements (%) of multivariate forecasting using GP according to mean absolute error (MAE) (dark red bars) and according to RMSE (dark blue bars), and improvements (%) of multivariate forecasting using an artificial neural network (ANN) according to MAE (orange bars) and according to RMSE (light blue bars) (PAIVA et al., 2020).

4.3 Specific Results

Complete results for each forecast horizon and location are presented in the Appendix. The most accurate results are in bold characters for both single and ensemble forecast comparisons. Model accuracy dominance depends on the location, forecast horizon and error metric, as summarized in Figure 4.4. The accomplished results point toward ANN as the most accurate for short horizons and GP as the most accurate for longer horizons, which also predominantly improves robustness. Furthermore, location attributes have been proven to affect model dominance. Figure 4.5 presents forecast accuracies for both methods applied to the Goiania station, where the most accurate results were obtained by ANN, and Figure 4.6 displays the results for the Desert Rock station, where the most accurate results were obtained by GP (PAIVA et al., 2020).

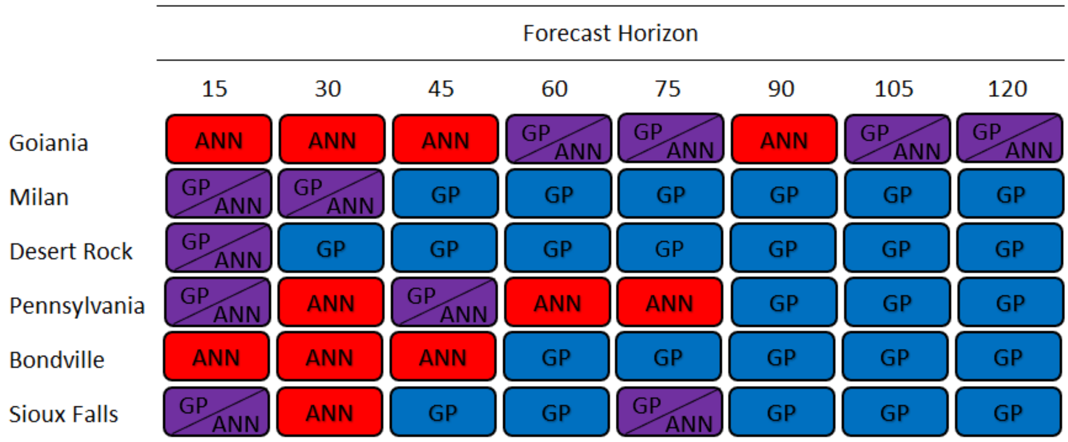


Figure 4.4 – Model accuracy dominance by location and forecast horizon in single forecasts. GP/ANN indicates cases in which accuracy dominance differs from the error metric evaluated (PAIVA et al., 2020).

Both GP and ANN methods were consistently improved considering both error metrics by employing an ensemble strategy for each forecast horizon and location. ANN presented more significant improvement and superior accuracy using the ensemble strategy in most cases, as summarized for model accuracy dominance in Figure 4.7 using ensemble forecasting. GP_{ens} led to the most accurate results in eight cases out of 48, while ANN_{ens} yielded the most accurate results in 23 cases out of 48. GP_{ens} achieved the most accurate results for the Milan station for horizons from 15 to 45 min and from 105 to 120 min using MAE as a reference metric. At Desert Rock station, GP_{ens} attained the lowest RMSE for horizons from 30 to 120 min. At Bondville station, GP_{ens} accomplished the lowest RMSE for horizons from 90 to 120 min and the lowest MAE for horizons from 75 to 120 min. At PSU station, GP_{ens} led to the lowest MAE and RMSE for horizons from 105 to 120 min. At Sioux Falls station, GP_{ens} yielded the lowest RMSE for horizons from 105 to 120 min and the lowest MAE for horizons of 45 min and from 90 to 120 min (PAIVA et al., 2020).

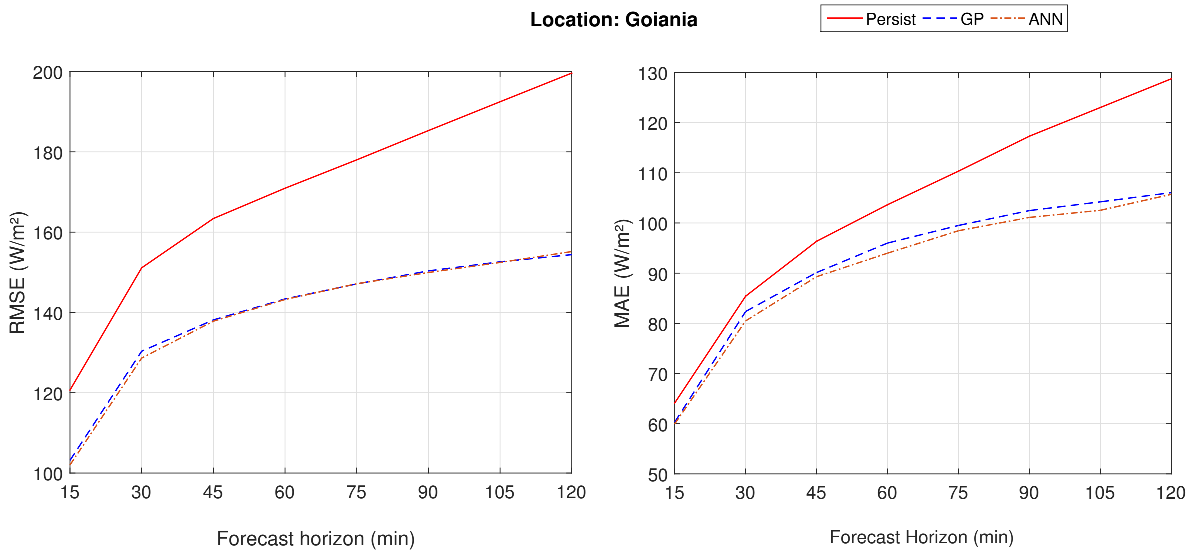


Figure 4.5 – Accuracy of persistence, GP and ANN according to RMSE (**left**) and MAE (**right**) for Goiania, showing the dominance of ANN (PAIVA et al., 2020).

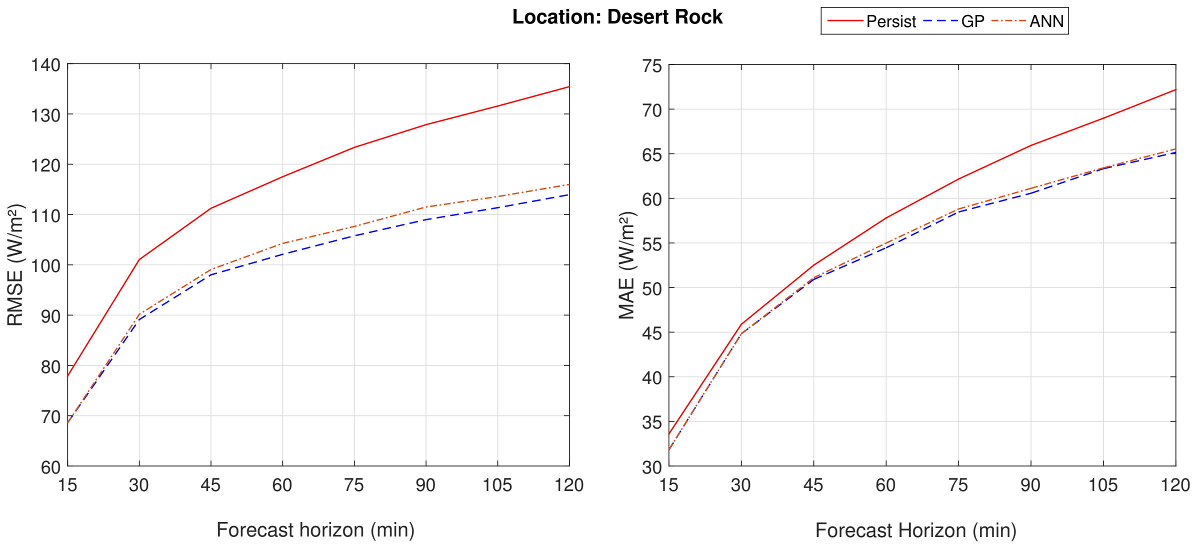


Figure 4.6 – Accuracy of persistence, GP and ANN according to RMSE (**left**) and MAE (**right**) for Desert Rock, showing the dominance of GP (PAIVA et al., 2020).

ANN_{ens} has proved to be consistently effective in the forecasts carried out for Goiania weather station, as expected, because the lower variations in sunshine duration along the year lead to a less biased dataset in terms of overfitting, as night period points are excluded from the dataset during the processing stage (PAIVA et al., 2020).

From a comparison of the results obtained by Haurwitz and Ineichen for clear sky index forecasts, it is possible to conclude that Ineichen k_t^* persistence produces lower errors than results obtained by Haurwitz for most of the locations and horizons of prediction.

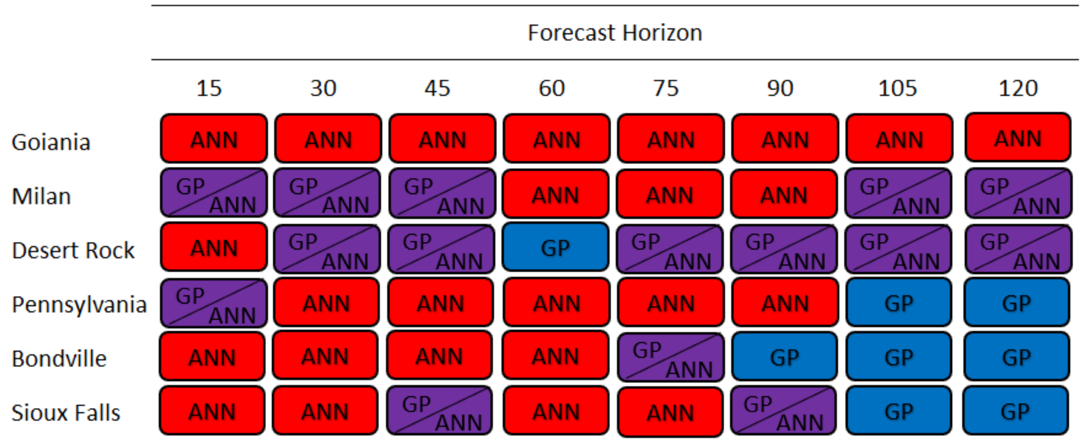


Figure 4.7 – Model accuracy dominance by location and forecast horizon in ensemble forecasts. GP/ANN indicates cases in which accuracy dominance differs from the error metric evaluated (PAIVA et al., 2020).

Nevertheless, as the AI methods used here are improved by employing exogenous inputs, a trend of clear sky model dominance over results from GP and ANN techniques was not achieved (PAIVA et al., 2020).

4.4 Generic Results

The computation of averages based on multiple results is widely employed as a procedure to achieve reliable generalized results according to Rana et al. (RANA; KOPRINSKA; AGELIDIS, 2016), although the use of averages does not disregard the importance of specific results. MAE and RMSE averages of all forecast horizons and locations were calculated in order to carry out a generic evaluation of accuracy for GP and ANN, and the results are presented in Figure 4.8. The average robustness of MAE and RMSE were similarly determined, and results are presented in Figure 4.9. From the generalized results, it is possible to assume that GP presents more accurate and robust forecast results in comparison to ANN for single forecasts; the ensemble strategy improves ANN forecasts more significantly than GP; the ANN ensemble generally presents the most accurate results; and both models produce similar forecastability, with little difference in terms of accuracy, indicating that GP can provide faster, more reliable and accurate predictions with lower computing complexity, while ANN can provide more accurate predictions using higher complexity and a time-demanding strategy (PAIVA et al., 2020).

A general comparison of clear sky indexes from multiple sites is exhibited in Table 4.1. From the analysis of results, it is possible to observe that the difference between Haurwitz k_t^* and Ineichen k_t^* forecast results is negligible, showing the low influence of the clear sky model on the accuracy of multivariate forecast results (PAIVA et al., 2020).

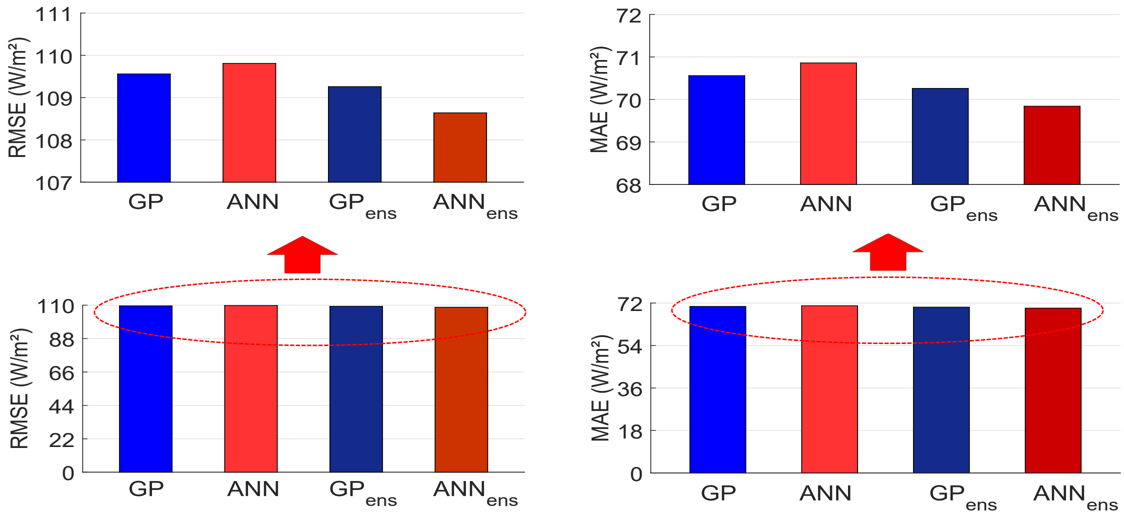


Figure 4.8 – General accuracies of GP, ANN, GP ensemble and ANN ensemble for all sites according to RMSE values (left-hand graphs) and MAE values (right-hand graphs) (PAIVA et al., 2020).

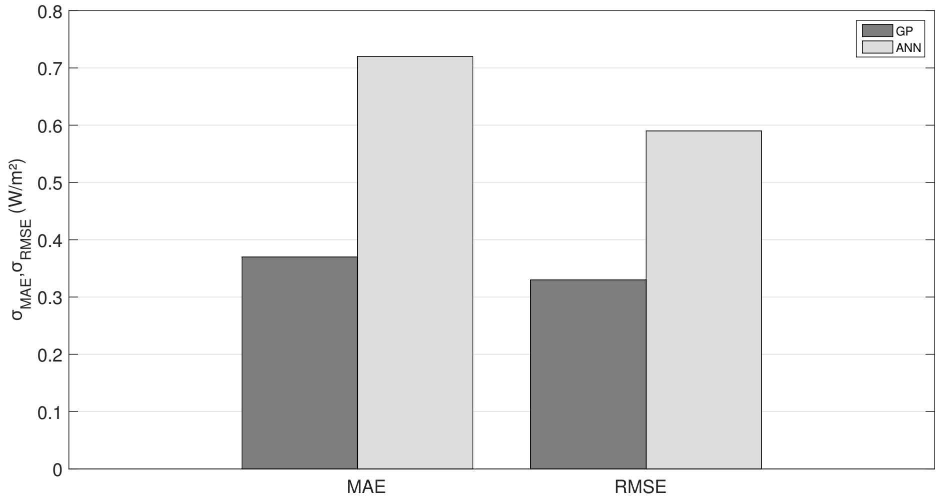


Figure 4.9 – Comparison of general robustness of GP and ANN single forecasts according to MAE and RMSE (PAIVA et al., 2020).

Table 4.1 – Generalized accuracies for Haurwitz k_t^* and Ineichen k_t^* forecasts (PAIVA et al., 2020).

	RMSE	$\boldsymbol{\sigma_{RMSE}}$	**MAE**	$\boldsymbol{\sigma_{MAE}}$
Haurwitz	111.87	0.44	70.22	0.54
Ineichen	111.93	0.47	70.33	0.55

4.5 Regression Functions

The following Equation presents an example of a regression function developed to forecast $\hat{k_t^*}(15)$, comprising a combination of the deterministic variable ω_s with previous values of k_t^* and the weather variables T_a and H_r. The algorithm has been proven to be efficient in selecting suitable variables to achieve accurate and robust models with generalization ability. Selected variables to develop regressions for Goiania station are expressed in Table 4.2 (PAIVA et al., 2020).

$$\hat{k_t^*}(15) = 0.535 + 0.98\tanh k_t^*(-5) - 0.0049[T_a(-45) + \omega_s \cdot k_t^*(-20)]$$
$$-0.142[e^{-k_t^*(-35)\cdot k_t^*(-50)} + \cos k_t^*(-20)] - 0.00141H_r(-5) + 0.0244[e^{-\omega_s}$$
$$-k_t^*(-5)k_t^*(-35)] + 0.00249\omega_s k_t^*(-20)e^{e^{-\omega_s}}$$

Table 4.2 – Variables selected by GP regression models according to the forecast horizon for Goiania (PAIVA et al., 2020).

Forecast Horizon	**Selected Variables**
15 min	ω_s, $k_t^*(-5,-20,-35,-50)$, $H_r(-5)$, $T_a(-40)$
30 min	t_s, ω_s, $\hat{k_t^*}(15)$, $k_t^*(-5)$, $p_a(-25)$, $H_r(-40)$
45 min	ω_s, $\hat{k_t^*}(30)$, $H_r(-5,-35,-40)$, $T_a(-20)$, $p_a(-60)$
60 min	t_s, h, $\hat{k_t^*}(15,45)$, $H_r(-15)$, $T_a(-40)$, $k_t^*(-45)$
75 min	ω_s, h, $\hat{k_t^*}(60)$, $p_a(-5,-10,-20)$, $T_a(-10,-55)$, $H_r(-10,-15)$, $W_s(-60)$
90 min	ω_s, $\hat{k_t^*}(30,45,75)$
105 min	ω_s, $\hat{k_t^*}(45,60,90)$
120 min	$Month$, $\hat{k_t^*}(105)$, $k_t^*(-25)$, $H_r(-30,-35)$, $T_a(-40)$

4.6 Comparison with the State-of-the-Art

A recent analysis of intraday solar irradiance forecasting at the SURFRAD weather stations has been carried out using regression and frequency domain models (REIKARD; HANSEN, 2019). A direct comparison of the results obtained by regression, frequency domain and MGGP is presented in Table 4.3. Reikard et al. (REIKARD; HANSEN, 2019) analyzed forecasts for the same years, based on the same historical data and datasets used here. Although pieces of datasets used in each analysis are not guaranteed to be the same, a direct comparison of the results is able to ensure the suitability of the results of GP prediction (PAIVA et al., 2020).

4.7 Machine Learning Algorithm Training Speed

Training machine learning algorithms to optimize results and accuracy is normally a time-consuming task. Table 4.4 presents a comparison of the average training times (in minutes) assessed for Goiania station according to each forecast horizon. Similar results were obtained for the other previously mentioned stations. Although MGGP has been demonstrated to be more robust for single forecasts, the training speed of this method is

Table 4.3 – Comparison of state-of-the-art methods applied to intraday solar irradiance forecasting for Surface Radiation Network (SURFRAD) weather stations (best values in bold) (PAIVA et al., 2020).

F.H.	Method	Desert Rock		Pennsylv. SU		Bondville		Sioux Falls	
		RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
	Regression	84.4	51.4	89.1	55.3	81.1	49.3	70.9	44.9
15	Freq. Domain	84.2	51.0	91.0	56.1	82.5	50.1	73.9	46.5
	GP_{ens}	**68.3**	**31.6**	**81.7**	**46.9**	**72.0**	**40.8**	**67.6**	**37.7**
	Regression	105.6	66.6	112.6	74.1	102.3	67.6	91.5	59.7
30	Freq. Domain	108.1	63.0	112.0	73.2	102.2	66.9	92.1	60.3
	GP_{ens}	**89.0**	**44.7**	**105.4**	**65.8**	**90.6**	**56.2**	**86.5**	**52.8**
	Regression	119.9	76.5	127.3	87.1	116.9	80.3	106.3	71.3
45	Freq. Domain	119.1	71.7	125.1	86.1	114.5	78.8	106.6	69.4
	GP_{ens}	**97.4**	**50.5**	**115.2**	**74.0**	**100.9**	**64.2**	**96.5**	**60.5**

lower than that for ANN. Improvements of MGGP parameter tuning strategies should be considered in future studies in order to increase the speed of MGGP training (PAIVA et al., 2020).

Table 4.4 – Comparison of training time required for each machine learning (ML) method, evaluated for the Goiania dataset (best values in bold) (PAIVA et al., 2020).

ML Method	F.H.							
	15	30	45	60	75	90	105	120
GP	3.62	3.36	3.24	3.40	3.50	3.71	3.43	3.42
ANN	**0.89**	**0.47**	**0.44**	**0.34**	**0.35**	**0.45**	**0.39**	**0.35**

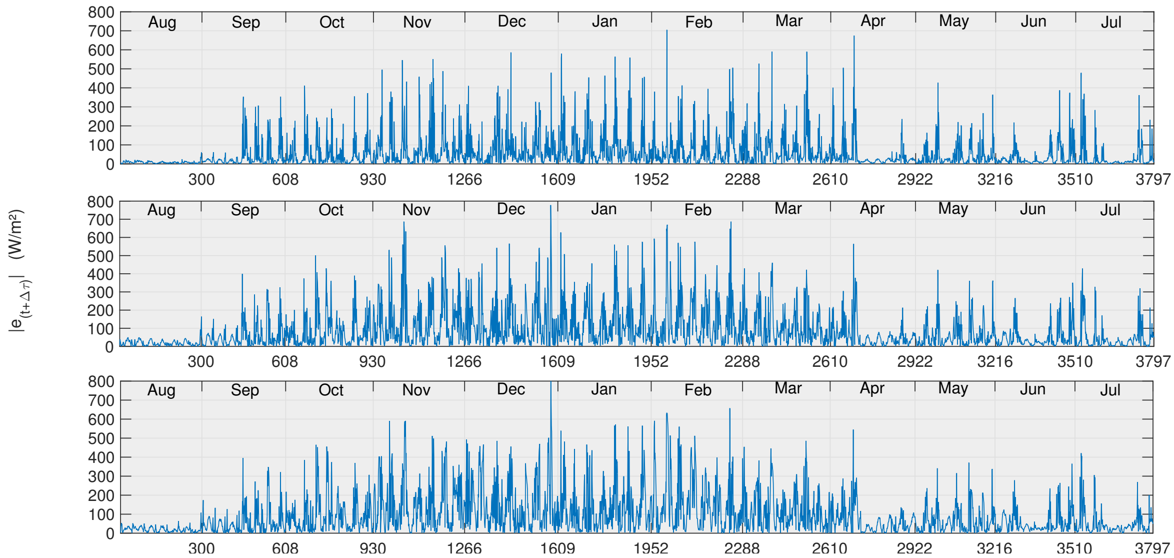
4.8 Analysis of individual and monthly errors

Analysis of individual and monthly errors and understanding site specific climate behavior may be useful for forecasting improvement. That is the case of analysis of Goiania forecasts. As discussed before, Goiania region is characterized by a dry winter season and a rainy summer (with higher solar variability). This aspect can be observed in individual errors and monthly errors plots.

Figure 4.10 presents plots of absolute individual errors for Goiania test dataset for forecast horizons of 15-min (top), 60-min (mid) and 120-min (bottom). It is possible to observe that errors are lower in the dry season (May–August)and higher in the rest of the year.

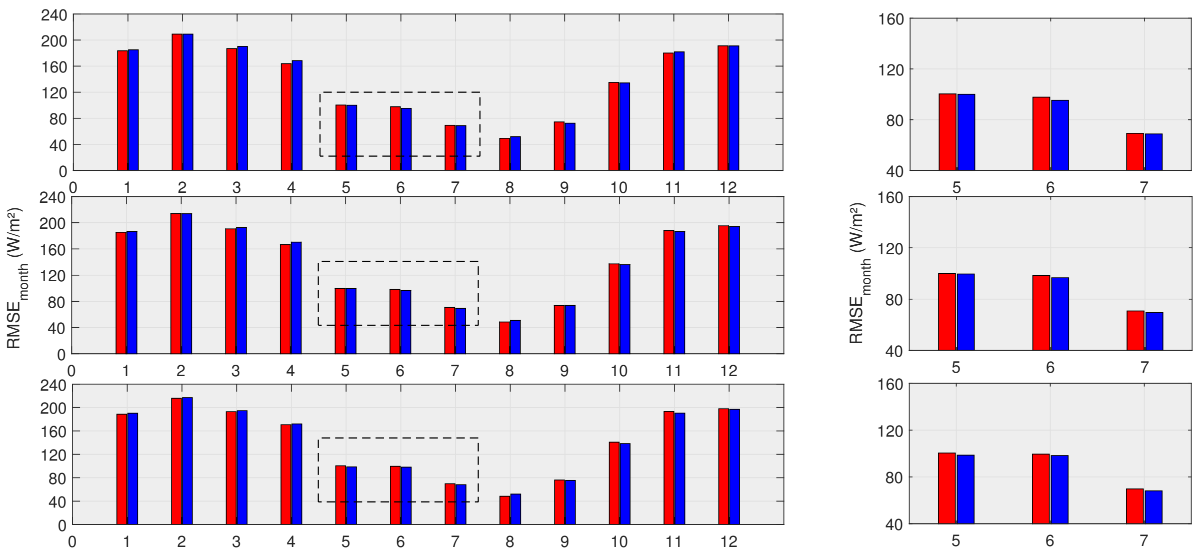
If the monthly RMSE value is calculated, it is observed that MGGP ensemble presents lower RMSE in May, June and July, even presenting an overall RMSE higher than ANN ensemble, as shows Figure 4.11. This may be explained by the fact that the algorithm minimizes the errors in the higher variability season and ends up with a worse performance in the dry season. This is an indicator for a possible successful hybridization strategy between the methods. In fact, literature has shown that combination of methods tends to improve overall accuracy.

Figure 4.10 – Individual errors plot from MGGP forecasts for Goiania test dataset for forecast horizons of 15-min (top), 60-min (mid) and 120-min (bottom).



.

Figure 4.11 – Monthly RMSE plots from $MGGP_{ens}$ (blue bars) and ANN_{ens} (red bars) forecasts for Goiania test dataset for forecast horizons of 90-min (top), 105-min (mid) and 120-min (bottom).



.

4.9 Summary

In this chapter, the results of MGGP tuning and forecasting have been presented and discussed. MGGP is a white-box method where derived regression functions have shown to build accurate and robust forecast models in a multivariate approach. It has shown competitive results in comparison with state-of-the-art MLP neural networks, particularly improving them for specific forecast horizons at some locations.

In the next chapter, the second case study is presented and its results are discussed: development of improved intraday PV power forecasting from a building integrated PV system installed at EMC–UFG based on MLP neural networks ensembles.

5 Case study II: Implementation and validation of multivariate intraday PV output power forecasting

This chapter presents results from tests of the second case covering intraday forecasting of PV output power from the EMC–UFG grid-connected PV system.

5.1 Combining meteorological station and PV power measurements

[1]As a second case study, this work evaluated applying of a physical hybrid artificial neural network model for intraday PV output power forecasting used to forecasting horizons of 15–120 minutes.

The forecast model combines weather variables measurements from the EMC–UFG meteorological station with PV output power measurements from a PV inverter of the EMC–UFG grid-connected PV system.

The complete PV system consists of a 34 kWp plant connected to the local grid with six single-phase inverters of 4.4 kW rated power and two single-phase inverters of 2.9 kW rated power. The rooftop presents a tilt angle of $\beta = 10°$. The azimuth angle γ_s of three 4.4 kW subsystems and two 2.9 kW face 15° West, and the other three 4.4 kW subsystems face 165° East. The analysis was carried out based on data of a subsystem facing 15° West. Figure 5.1 presents a google earth view of the EMC–UFG grid connected PV system, showing PV modules with 15° West azimuth and 165° East azimuth angles.

PV power measurements were obtained at inverter DC side, within the period from April 5th, 2017 to November 11th, 2017 from a 4.4 kW inverter manufactured by Eltek®, which is connected to 20 polycrystalline PV panels of 235 Wp manufactured by SunEarth®, and The Weather data were obtained for the same period as the PV values. DC PV power data were used instead of AC PV power because it is not influenced by inverter efficiency.

5.2 Data processing

The proposed methodology shown in Figure 3.1 was adapted to PV output power forecasting. After data quality control and removal of PV measurements below $h = 5°$, the PV output power (DC) was first normalized by the PV system capacity, Pnom = 4700 Wp, creating a virtual 1 kWp PV generator. After that, a new definition of clear sky index

[1] Some texts from this session onwards have been extracted from this research publication version at *IEEE Powertech 2019* conference (Paiva et al., 2019).

Figure 5.1 – Google earth view of the EMC–UFG grid connected PV system.

Reference: authors.

was developed, following the k_{PV} index concept from Engerer e Mills (2014). The k_{PV} is obtained by the following Equation:

$$k_{PV} = \frac{P_o}{G_{clr,T}} \tag{5.1}$$

Where P_o is the PV output power in a 15-min time window and $G_{clr,T}$ is a transposed clear sky irradiance for a PV module tilt angle β and azimuth angle γ_s. $G_{clr,T}$ was obtained by the Reindl model as it yielded good results under clear days (ENGERER; MILLS, 2014). Equation 5.2 presents formulation to $G_{clr,T}$.

$$G_{clr,T} = (G_b + G_d A_i)R_b + G_d(1 - A_i)(\frac{1 + \cos\beta}{2})[1 + f\sin^3\frac{\beta}{2}] + G\rho_g(\frac{1 - \cos\beta}{2}) \tag{5.2}$$

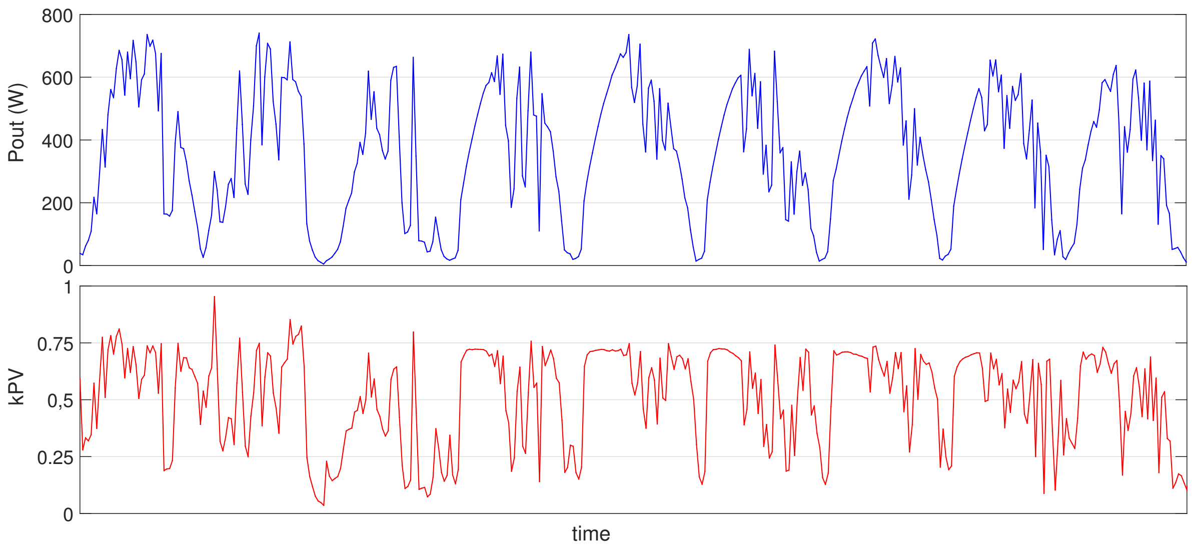
Where G is the Ineichen clear sky model, given by Equation 3.4. Diffuse G_d and beam G_b components of the clear sky irradiance and other parameters in Equation 5.2 were obtained following algorithms described by Duffie e Beckman (2013).

Figure 5.2 shows the normalized index distributed over some days of the dataset. It is possible to observe that the k_{PV} index successfully reduces the seasonality in PV output power measurements.

5.3 Input-output scheme

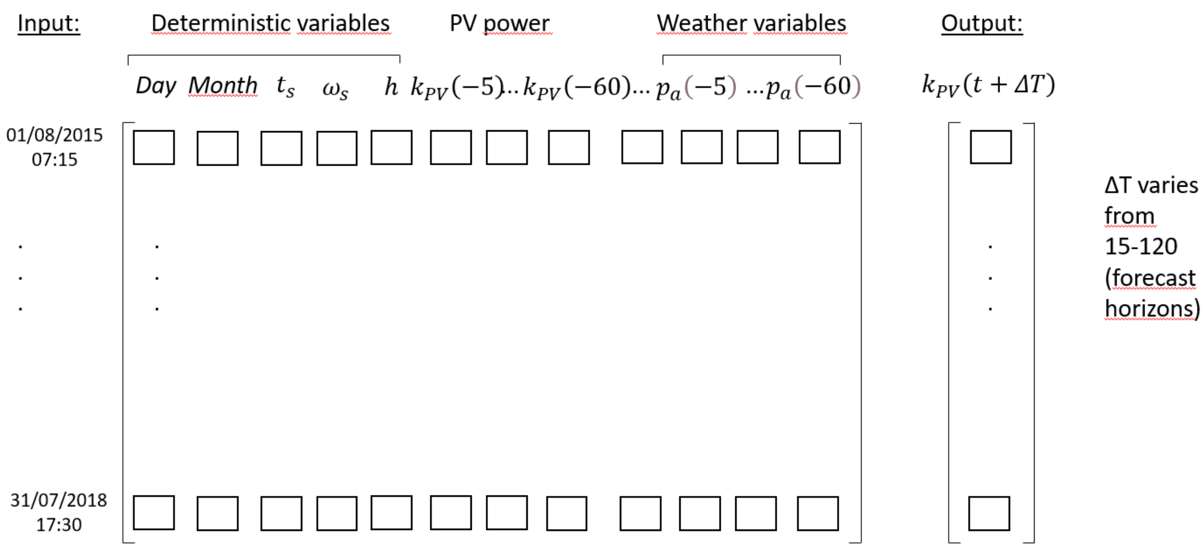
The input-output scheme for the PV power forecast is presented in Figure 5.3. In this case, each forecast horizon forecast used a total of 77 inputs.

Figure 5.2 – Virtual 1 kWp PV plant output power (top) and k_{PV} distribution (bottom).



Reference: (Paiva et al., 2019).

Figure 5.3 – Data relations of input and output vectors for the case of PV output power forecast model developed.



5.4 Forecasting results

In this case, 20 neural networks were trained to build MLP Neural Networks ensembles. The PV output power forecasting results from the MLP Neural Networks were validated by comparing with the Persistence of the k_{PV} index. The results are presented in Figure 5.4.

Figure 5.4 – Table of comparison of the forecast results from ANN ensemble versus a k_{PV} persistence from Paiva et al. (2019) (best results presented in bold) .

	MAE (W/m^2)		*RMSE (W/m^2)*		*EMAE*		*OMAE*		*s*
Forecast Horizon	*Persist*	*ANN*	*Persist*	*ANN*	*Persist*	*ANN*	*Persist*	*ANN*	*ANN*
15 min	**33.65**	34.17	66.61	**60.08**	**9.22**	9.30	**5.30**	5.38	**9.80**
30 min	46.37	**43.19**	81.87	**73.17**	12.52	**11.57**	7.30	**6.80**	**10.63**
45 min	55.00	**48.29**	93.29	**81.80**	14.71	**12.82**	8.66	**7.60**	**12.32**
60 min	61.23	**51.87**	100.78	**87.50**	16.29	**13.68**	9.64	**8.17**	**13.18**
75 min	67.67	**54.87**	108.95	**90.40**	17.90	**14.46**	10.65	**8.64**	**17.03**
90 min	73.53	**57.57**	115.88	**91.31**	19.35	**15.14**	11.58	**9.06**	**21.20**
105 min	77.93	**60.48**	122.47	**94.36**	20.43	**15.86**	12.27	**9.52**	**22.96**
120 min	83.83	**62.52**	131.12	**99.98**	21.86	**16.25**	13.20	**9.84**	**23.75**

The results show that forecast skills range from 9.79% to 23.75%. It stands for ANN ensemble consistently improves the smart persistence model for the PV output power forecast.

5.5 Summary

This chapter described the MGGP results and discussions for tuning and solar forecasting. MGGP is a white-box method whose derived regression functions built accurate and robust forecast models in a multivariate approach. It has presented competitive results compared to state-of-the-art MLP neural networks, mainly improving forecast horizons at some locations.

The next chapter regards the second case study, presenting and discussing improved intraday PV power forecasting results from a building-integrated PV system installed at EMC–UFG based on MLP neural network ensembles.

6 Conclusions

Machine learning algorithms are extensively adopted techniques for solar forecasting. This research proposed and evaluated multigene genetic programming (MGGP) as a novel machine learning algorithm classified as a white box to perform intraday solar irradiance forecasting. MGGP derives analytical regression functions that can be implemented without a specific software environment in any modern programming language using basic hardware. MGGP proved to consistently possess data generalization ability, providing robust and reliable solutions. The MGGP algorithm and another state-of-the-art MLP artificial neural network (ANN) algorithm were applied to datasets from six locations from three countries to compare results for forecast horizons from 15 to 120 minutes.

Data processing strategies were carefully analyzed in terms of input and output alternatives. Initial simulations were carried out for solar irradiance forecasting, using fifteen-minute time-windows as input data. Five-minute time-window data, Haurwitz and Ineichen clear sky indexes were considered and combined with solar deterministic variables and weather variables to improve forecast accuracy in terms of the data processing strategy.

The computation of MAE and RMSE as error metrics showed that the location, forecast horizon and error of evaluation influence the selection of the dominant model in terms of accuracy. MGGP and ANN typically yielded similar and consistent results. MGGP utilization for single forecasts led to more accurate and robust results as opposed to ANN. Predictions were significantly improved for MGGP and ANN by adopting ensemble forecast, while the ensemble strategy improved ANN more extensively than MGGP. Regarding ensemble forecasts, MGGP was more accurate for a lower number of locations and evaluated forecast horizons than ANN, presenting the best forecast skills for Desert Rock station. MGGP predominantly accomplished more accurate prediction results for longer forecast horizons from 90 to 120 minutes ahead for different localities.

Based on a direct comparison with other state-of-the-art forecasting methods applied to the same locations in the USA, MGGP presented a relevant reduction in error and proved to be a reliable and accurate approach for the analyzed localities. Besides, locality attributes demonstrated to affect model dominance, indicating that both MGGP and ANN are suitable to diverse locations.

As a suggestion, future studies may address hybridization strategies, ML algorithm enhancements, advanced data processing strategies applied to MGGP forecasting, and improvements in parameter tuning to enhance MGGP's training speed. Moreover, the inclusion of other solar parameters may be studied to improve solar forecasting techniques' accuracy and performance.

A PV output power forecasting model based on MLP neural network ensemble yielded promising results. Future works may compare the proposed methodology with forecasting models based only on a reference meteorological station. Other statistical methods and hybridization strategies may be investigated for PV output power forecasting.

Bibliography

AGGARWAL, S.; SAINI, L. Solar energy prediction using linear and non-linear regularization models: A study on ams (american meteorological society) 2013–14 solar energy prediction contest. *Energy*, v. 78, p. 247 – 256, 2014. ISSN 0360-5442. Disponível em: <http://www.sciencedirect.com/science/article/pii/S036054421401161X>. Citado na página 44.

AGUIAR, L. M. et al. Combining solar irradiance measurements, satellite-derived data and a numerical weather prediction model to improve intra-day solar forecasting. *Renewable Energy*, v. 97, p. 599 – 610, 2016. ISSN 0960-1481. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0960148116305390>. Citado na página 24.

ALMEIDA, M. P. et al. Comparative study of pv power forecast using parametric and nonparametric pv models. *Solar Energy*, v. 155, p. 854 – 866, 2017. ISSN 0038-092X. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0038092X17306175>. Citado 2 vezes nas páginas 34 and 44.

Alobaidi, M. H. et al. Mapping of the solar irradiance in the uae using advanced artificial neural network ensemble. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, v. 7, n. 8, p. 3668–3680, Aug 2014. ISSN 2151-1535. Citado na página 44.

Bakirtzis, E. A.; Biskas, P. N. Multiple time resolution stochastic scheduling for systems with high renewable penetration. *IEEE Transactions on Power Systems*, v. 32, n. 2, p. 1030–1040, 2017. Citado na página 23.

Barcelo, W. R.; Rastgoufard, P. Dynamic economic dispatch using the extended security constrained economic dispatch algorithm. *IEEE Transactions on Power Systems*, v. 12, n. 2, p. 961–967, 1997. Citado na página 23.

Ben Taieb, S. et al. A review and comparison of strategies for multi-step ahead time series forecasting based on the nn5 forecasting competition. *Expert Systems with Applications*, v. 39, n. 8, p. 7067 – 7083, 2012. ISSN 0957-4174. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0957417412000528>. Citado na página 44.

BENALI, L. et al. Solar radiation forecasting using artificial neural network and random forest methods: Application to normal beam, horizontal diffuse and global components. *Renewable Energy*, v. 132, p. 871 – 884, 2019. ISSN 0960-1481. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0960148118309947>. Citado na página 42.

Bessa, R. J.; Trindade, A.; Miranda, V. Spatial-temporal solar power forecasting for smart grids. *IEEE Transactions on Industrial Informatics*, v. 11, n. 1, p. 232–241, 2015. Citado na página 38.

BILIONIS, I.; CONSTANTINESCU, E. M.; ANITESCU, M. Data-driven model for solar irradiation based on satellite observations. *Solar Energy*, v. 110, p. 22 – 38, 2014. ISSN 0038-092X. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0038092X14004393>. Citado na página 44.

BOUZERDOUM, M.; MELLIT, A.; Massi Pavan, A. A hybrid model (sarima–svm) for short-term power forecasting of a small-scale grid-connected photovoltaic plant. *Solar Energy*, v. 98, p. 226 – 235, 2013. ISSN 0038-092X. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0038092X13004039>. Citado na página 44.

BOX, G. E. et al. *Time series analysis: forecasting and control*. [S.l.]: John Wiley & Sons, 2015. Citado na página 38.

BREIMAN, L. Bagging predictors. *Machine learning*, Springer, v. 24, n. 2, p. 123–140, 1996. Citado na página 42.

BROWNLEE, J. *Clever Algorithms: Nature-Inspired Programming Recipes*. 1st. ed. [S.l.]: Lulu.com, 2011. ISBN 1446785068, 9781446785065. Citado na página 61.

BURROWS, W. R. Cart regression models for predicting uv radiation at the ground in the presence of cloud and other environmental factors. *Journal of Applied Meteorology*, American Meteorological Society, Boston MA, USA, v. 36, n. 5, p. 531 – 544, 1997. Disponível em: <https://journals.ametsoc.org/view/journals/apme/36/5/1520-0450_1997_036_0531_crmfpu_2.0.co_2.xml>. Citado na página 44.

CAO, J.; LIN, X. Study of hourly and daily solar irradiation forecast using diagonal recurrent wavelet neural networks. *Energy Conversion and Management*, v. 49, n. 6, p. 1396 – 1406, 2008. ISSN 0196-8904. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0196890408000125>. Citado na página 44.

CHAKRABORTY, P. et al. Fine-grained photovoltaic output prediction using a bayesian ensemble. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. [S.l.: s.n.], 2012. v. 26, n. 1. Citado na página 44.

CHU, Y.; PEDRO, H. T.; COIMBRA, C. F. Hybrid intra-hour dni forecasts with sky image processing enhanced by stochastic learning. *Solar Energy*, v. 98, p. 592 – 603, 2013. ISSN 0038-092X. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0038092X13004325>. Citado na página 44.

CHU, Y. et al. Real-time forecasting of solar irradiance ramps with smart image processing. *Solar Energy*, v. 114, p. 91 – 104, 2015. ISSN 0038-092X. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0038092X15000389>. Citado na página 44.

CORTES, C.; VAPNIK, V. Support-vector networks. *Machine learning*, Springer, v. 20, n. 3, p. 273–297, 1995. Citado 2 vezes nas páginas 9 and 41.

De Felice, M.; PETITTA, M.; RUTI, P. M. Short-term predictability of photovoltaic production over italy. *Renewable Energy*, v. 80, p. 197 – 204, 2015. ISSN 0960-1481. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0960148115001007>. Citado na página 44.

de Lima, E. B. et al. Tuning genetic programming parameters with factorial designs. In: *IEEE Congress on Evolutionary Computation*. [S.l.: s.n.], 2010. p. 1–8. ISSN 1089-778X. Citado na página 67.

de Paiva, G. M. et al. Assessment of exogenous variables on intra-day solar irradiance forecasting models. In: *2018 IEEE International Conference on Environment and Electrical Engineering and 2018 IEEE Industrial and Commercial Power Systems Europe (EEEIC / I CPS Europe)*. [S.l.: s.n.], 2018. p. 1–6. Citado na página 19.

de Paiva, G. M. et al. Intelligent approach to improve genetic programming based intra-day solar forecasting models. In: *2018 IEEE Congress on Evolutionary Computation (CEC)*. [S.l.: s.n.], 2018. p. 1–8. Citado 2 vezes nas páginas 19 and 54.

Demirtas, M. et al. Prediction of solar radiation using meteorological data. In: *2012 International Conference on Renewable Energy Research and Applications (ICRERA)*. [S.l.: s.n.], 2012. p. 1–4. Citado na página 44.

DIAGNE, M. et al. Review of solar irradiance forecasting methods and a proposition for small-scale insular grids. *Renewable and Sustainable Energy Reviews*, v. 27, p. 65 – 76, 2013. ISSN 1364-0321. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1364032113004334>. Citado 5 vezes nas páginas 9, 28, 29, 30, and 31.

DOLARA, A.; LEVA, S.; MANZOLINI, G. Comparison of different physical models for pv power output prediction. *Solar Energy*, v. 119, p. 83 – 99, 2015. ISSN 0038-092X. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0038092X15003254>. Citado na página 34.

DONG, Z. et al. A novel hybrid approach based on self-organizing maps, support vector regression and particle swarm optimization to forecast solar irradiance. *Energy*, v. 82, p. 570 – 577, 2015. ISSN 0360-5442. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0360544215000900>. Citado na página 44.

DUFFIE, J. A.; BECKMAN, W. A. *Solar Engineering of Thermal Processes*. [S.l.]: John Wiley & Sons, Inc., Hoboken, New Jersey, 2013. Citado 3 vezes nas páginas 26, 55, and 79.

EEA. *Electricity production by fuel*. 2012. Website. European Environment Agency. Disponível em: <https://www.eea.europa.eu/data-and-maps/indicators/electricity-production-by-fuel-1/electricity-production-by-fuel-assessment-3>. Citado na página 23.

EIA. *Electricity in the United States*. 2020. Website. U.S. Energy Information Administration. Disponível em: <https://www.eia.gov/energyexplained/electricity/electricity-in-the-us.php>. Citado na página 23.

EIBEN, A.; SMIT, S. Parameter tuning for configuring and analyzing evolutionary algorithms. *Swarm and Evolutionary Computation*, v. 1, n. 1, p. 19 – 31, 2011. ISSN 2210-6502. Disponível em: <http://www.sciencedirect.com/science/article/pii/S2210650211000022>. Citado na página 67.

Ela, E.; O'Malley, M. Studying the variability and uncertainty impacts of variable generation at multiple timescales. *IEEE Transactions on Power Systems*, v. 27, n. 3, p. 1324–1333, 2012. Citado na página 23.

ELMAN, J. L. Finding structure in time. *Cognitive science*, Wiley Online Library, v. 14, n. 2, p. 179–211, 1990. Citado na página 39.

ENGERER, N.; MILLS, F. Kpv: A clear-sky index for photovoltaics. *Solar Energy*, v. 105, p. 679 – 693, 2014. ISSN 0038-092X. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0038092X14002151>. Citado 3 vezes nas páginas 37, 56, and 79.

EPE. *Brazilian Energy Balance Year 2008*. [S.l.], 2009. Disponível em: <www.epe.gov.br>. Citado na página 23.

EPE. *Balanço Energético Nacional 2019*. [S.l.], 2019. Ano Base 2018. Disponível em: <www.epe.gov.br>. Citado na página 23.

EUROSTAT. *What is the source of electricity we consume?* 2020. Website. Disponível em: <https://ec.europa.eu/eurostat>. Citado na página 23.

FERNÁNDEZ, Á.; GALA, Y.; DORRONSORO, J. R. Machine learning prediction of large area photovoltaic energy production. In: WOON, W. L.; AUNG, Z.; MADNICK, S. (Ed.). *Data Analytics for Renewable Energy Integration*. Cham: Springer International Publishing, 2014. p. 38–53. ISBN 978-3-319-13290-7. Citado na página 44.

Ferrari, S. et al. Illuminance prediction through extreme learning machines. In: *2012 IEEE Workshop on Environmental Energy and Structural Monitoring Systems (EESMS)*. [S.l.: s.n.], 2012. p. 97–103. Citado na página 44.

GARG, A.; SRIRAM, S.; TAI, K. Empirical analysis of model selection criteria for genetic programming in modeling of time series system. In: *2013 IEEE Conference on Computational Intelligence for Financial Engineering Economics (CIFEr)*. [S.l.: s.n.], 2013. p. 90–94. ISSN 2380-8454. Citado na página 61.

GASTÓN, M. et al. A new Adaptive methodology of Global-to-Direct irradiance based on clustering and kernel machines techniques. In: *15th SolarPACES Conference*. Berlin, Germany: [s.n.], 2010. p. 11693. Disponível em: <https://hal.archives-ouvertes.fr/hal-00919064>. Citado na página 44.

GHIMIRE, S. et al. Global solar radiation prediction by ann integrated with european centre for medium range weather forecast fields in solar rich cities of queensland australia. *Journal of Cleaner Production*, v. 216, p. 288 – 310, 2019. ISSN 0959-6526. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0959652619301775>. Citado na página 61.

GHORBANI, M. A. et al. Chaos-based multigene genetic programming: A new hybrid strategy for river flow forecasting. *Journal of Hydrology*, v. 562, p. 455 – 467, 2018. ISSN 0022-1694. Disponível em: <http://www.sciencedirect.com/science/article/pii/S002216941830307X>. Citado na página 61.

GUIDE, M. U. Neural network toolbox. *The MathWorks*, 2002. Citado na página 40.

GUTIERREZ-COREA, F.-V. et al. Forecasting short-term solar irradiance based on artificial neural networks and data from neighboring meteorological stations. *Solar Energy*, v. 134, p. 119 – 131, 2016. ISSN 0038-092X. Citado 2 vezes nas páginas 32 and 33.

INEICHEN, P. Comparison of eight clear sky broadband models against 16 independent data banks. *Solar Energy*, v. 80, n. 4, p. 468 – 478, 2006. ISSN 0038-092X. Urban Ventilation. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0038092X05001635>. Citado na página 56.

INEICHEN, P.; PEREZ, R. A new airmass independent formulation for the linke turbidity coefficient. *Solar Energy*, v. 73, n. 3, p. 151 – 157, 2002. ISSN 0038-092X. Citado na página 55.

INMAN, R. H.; PEDRO, H. T.; COIMBRA, C. F. Solar forecasting methods for renewable energy integration. *Progress in Energy and Combustion Science*, v. 39, n. 6, p. 535 – 576, 2013. ISSN 0360-1285. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0360128513000294>. Citado 3 vezes nas páginas 9, 28, and 29.

JACOBSON, M. Z. Short-term effects of controlling fossil-fuel soot, biofuel soot and gases, and methane on climate, arctic ice, and air pollution health. *Journal of Geophysical Research: Atmospheres*, v. 115, n. D14, 2010. Disponível em: <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2009JD013795>. Citado na página 23.

KALLIO-MYERS, V. et al. Global horizontal irradiance forecast for finland based on geostationary weather satellite data. *Solar Energy*, v. 198, p. 68 – 80, 2020. ISSN 0038-092X. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0038092X20300141>. Citado na página 24.

KLEISSL, J. *Solar energy forecasting and resource assessment*. [S.l.]: Academic Press, 2013. Citado na página 31.

KOZA, J. R. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. Cambridge, MA, USA: MIT Press, 1992. ISBN 0-262-11170-5. Citado na página 61.

Kundur, P. et al. Definition and classification of power system stability ieee/cigre joint task force on stability terms and definitions. *IEEE Transactions on Power Systems*, v. 19, n. 3, p. 1387–1401, 2004. Citado na página 23.

LAURET, P. et al. A benchmarking of machine learning techniques for solar radiation forecasting in an insular context. *Solar Energy*, v. 112, p. 446 – 457, 2015. ISSN 0038-092X. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0038092X14006057>. Citado na página 44.

LAZZARONI, M. et al. Models for solar radiation prediction based on different measurement sites. *Measurement*, v. 63, p. 346 – 363, 2015. ISSN 0263-2241. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0263224114006162>. Citado na página 44.

LEE, Y.-S.; TONG, L.-I. Forecasting time series using a methodology based on autoregressive integrated moving average and genetic programming. *Knowledge-Based Systems*, v. 24, n. 1, p. 66 – 72, 2011. ISSN 0950-7051. Disponível em:

<http://www.sciencedirect.com/science/article/pii/S0950705110001127>. Citado na página 61.

LEVA, S. et al. Analysis and validation of 24 hours ahead neural network forecasting of photovoltaic output power. *Mathematics and Computers in Simulation*, v. 131, p. 88 – 100, 2017. ISSN 0378-4754. 11th International Conference on Modeling and Simulation of Electric Machines, Converters and Systems. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0378475415001238>. Citado 5 vezes nas páginas 9, 45, 46, 64, and 65.

Leva, S. et al. Pv plant power nowcasting: A real case comparative study with an open access dataset. *IEEE Access*, v. 8, p. 194428–194440, 2020. Citado na página 47.

LI, Y.; SU, Y.; SHU, L. An armax model for forecasting the power output of a grid connected photovoltaic system. *Renewable Energy*, v. 66, p. 78 – 89, 2014. ISSN 0960-1481. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0960148113006551>. Citado na página 39.

LIMA, F. J. et al. Forecast for surface solar irradiance at the brazilian northeastern region using nwp model and artificial neural networks. *Renewable Energy*, v. 87, p. 807 – 818, 2016. ISSN 0960-1481. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0960148115304249>. Citado na página 33.

Lorenz, E. et al. Irradiance forecasting for the power prediction of grid-connected photovoltaic systems. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, v. 2, n. 1, p. 2–10, 2009. Citado na página 33.

MARQUEZ, R.; COIMBRA, C. F. Forecasting of global and direct solar irradiance using stochastic learning methods, ground experiments and the nws database. *Solar Energy*, v. 85, n. 5, p. 746 – 756, 2011. ISSN 0038-092X. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0038092X11000193>. Citado na página 44.

MARQUEZ, R.; PEDRO, H. T.; COIMBRA, C. F. Hybrid solar forecasting method uses satellite imaging and ground telemetry as inputs to anns. *Solar Energy*, v. 92, p. 176 – 188, 2013. ISSN 0038-092X. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0038092X13000881>. Citado 2 vezes nas páginas 34 and 35.

MCGOVERN, A. et al. Solar energy prediction: An international contest to initiate interdisciplinary research on compelling meteorological problems. *Bulletin of the American Meteorological Society*, American Meteorological Society, Boston MA, USA, v. 96, n. 8, p. 1388 – 1395, 2015. Disponível em: <https://journals.ametsoc.org/view/journals/bams/96/8/bams-d-14-00006.1.xml>. Citado na página 44.

MEHR, A. D.; JABARNEJAD, M.; NOURANI, V. Pareto-optimal mpsa-mggp: A new gene-annealing model for monthly rainfall forecasting. *Journal of Hydrology*, v. 571, p. 406 – 415, 2019. ISSN 0022-1694. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0022169419301362>. Citado na página 61.

MEHR, A. D.; KAHYA, E.; OLYAIE, E. Streamflow prediction using linear genetic programming in comparison with a neuro-wavelet technique. *Journal of Hydrology*, v. 505, p. 240 – 249, 2013. ISSN 0022-1694. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0022169413007105>. Citado na página 61.

MESSAGIE, M. et al. The hourly life cycle carbon footprint of electricity generation in belgium, bringing a temporal resolution in life cycle assessment. *Applied Energy*, v. 134, p. 469 – 476, 2014. ISSN 0306-2619. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0306261914008824>. Citado na página 23.

MIHALAKAKOU, G.; SANTAMOURIS, M.; ASIMAKOPOULOS, D. The total solar radiation time series simulation in athens, using neural networks. *Theoretical and Applied Climatology*, v. 66, n. 3-4, p. 185–197, 2000. Cited By 65. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-0033861306&doi=10.1007%2fs007040070024&partnerID=40&md5=2c02e05f56cc0181f3fb900c09896fe6>. Citado na página 44.

MILLER, S. D. et al. Short-term solar irradiance forecasting via satellite/model coupling. *Solar Energy*, v. 168, p. 102 – 117, 2018. ISSN 0038-092X. Advances in Solar Resource Assessment and Forecasting. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0038092X17310435>. Citado na página 35.

MORENO, A.; GILABERT, M.; MARTíNEZ, B. Mapping daily global solar irradiation over spain: A comparative study of selected approaches. *Solar Energy*, v. 85, n. 9, p. 2072 – 2084, 2011. ISSN 0038-092X. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0038092X11001976>. Citado na página 44.

Mori, H.; Kosemura, N. Optimal regression tree based rule discovery for short-term load forecasting. In: *2001 IEEE Power Engineering Society Winter Meeting. Conference Proceedings (Cat. No.01CH37194)*. [S.l.: s.n.], 2001. v. 2, p. 421–426 vol.2. Citado na página 44.

Mori, H.; Takahashi, A. A data mining method for selecting input variables for forecasting model of global solar radiation. In: *PES T D 2012*. [S.l.: s.n.], 2012. p. 1–6. ISSN 2160-8563. Citado na página 44.

NOBRE, A. M. et al. Pv power conversion and short-term forecasting in a tropical, densely-built environment in singapore. *Renewable Energy*, v. 94, p. 496 – 509, 2016. ISSN 0960-1481. Citado 4 vezes nas páginas 12, 31, 32, and 39.

OGLIARI, E. et al. Physical and hybrid methods comparison for the day ahead pv output power forecast. *Renewable Energy*, v. 113, p. 11 – 21, 2017. ISSN 0960-1481. Disponível em: <http://www.sciencedirect.com/science/article/pii/S096014811730455X>. Citado 3 vezes nas páginas 9, 33, and 34.

OLAIYA, F.; ADEYEMO, A. B. Application of data mining techniques in weather prediction and climate change studies. *International Journal of Information Engineering and Electronic Business*, Modern Education and Computer Science Press, v. 4, n. 1, p. 51, 2012. Citado na página 44.

PAIVA, G. M. et al. Analysis of inverter sizing ratio for pv systems considering local climate data in central brazil. *IET Renewable Power Generation*, v. 11, n. 11, p. 1364–1370, 2017. ISSN 1752-1416. Citado na página 34.

Paiva, G. M. d. et al. Intra-day forecasting of building-integrated pv systems for power systems operation using ann ensemble. In: *2019 IEEE Milan PowerTech*. [S.l.: s.n.], 2019. p. 1–5. Citado 6 vezes nas páginas 11, 19, 47, 78, 80, and 81.

PAIVA, G. Mendonça de et al. Multiple site intraday solar irradiance forecasting by machine learning algorithms: Mggp and mlp neural networks. *Energies*, v. 13, n. 11, 2020. ISSN 1996-1073. Disponível em: <https://www.mdpi.com/1996-1073/13/11/3005>. Citado 24 vezes nas páginas 10, 11, 12, 19, 49, 54, 55, 56, 59, 60, 61, 62, 63, 64, 65, 67, 68, 69, 70, 71, 72, 73, 74, and 75.

PAIVA SéRGIO PIRES PIMENTEL, B. P. d. A. E. G. M. Gabriel Mendonça de. RegressÃo simbÓlica aplicada na previsÃo de irradiÂncia solar intra-diÁria na cidade de goiÂnia (brasil). *Revista Brasileira de Energia Solar*, v. 9, n. 2, p. 9, 2018. ISSN 2526-2831. Disponível em: <https://rbens.emnuvens.com.br/rbens/article/view/247>. Citado na página 19.

PAN, I.; PANDEY, D. S.; DAS, S. Global solar irradiation prediction using a multi-gene genetic programming approach. *Journal of Renewable and Sustainable Energy*, v. 5, n. 6, p. 063129, 2013. Disponível em: <https://doi.org/10.1063/1.4850495>. Citado na página 61.

PAOLI, C. et al. Forecasting of preprocessed daily solar radiation time series using neural networks. *Solar Energy*, v. 84, n. 12, p. 2146 – 2160, 2010. ISSN 0038-092X. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0038092X10002793>. Citado na página 44.

PEDRO, H. T.; COIMBRA, C. F. Assessment of forecasting techniques for solar power production with no exogenous inputs. *Solar Energy*, v. 86, n. 7, p. 2017 – 2028, 2012. ISSN 0038-092X. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0038092X12001429>. Citado na página 42.

PEDRO, H. T.; COIMBRA, C. F. Nearest-neighbor methodology for prediction of intra-hour global horizontal and direct normal irradiances. *Renewable Energy*, v. 80, p. 770 – 782, 2015. ISSN 0960-1481. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0960148115001792>. Citado na página 44.

PEDRO, H. T.; COIMBRA, C. F. Short-term irradiance forecastability for various solar micro-climates. *Solar Energy*, v. 122, p. 587 – 602, 2015. ISSN 0038-092X. Citado 5 vezes nas páginas 12, 31, 32, 37, and 64.

PEDRO, H. T. et al. Assessment of machine learning techniques for deterministic and probabilistic intra-hour solar forecasts. *Renewable Energy*, v. 123, p. 191 – 203, 2018. ISSN 0960-1481. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0960148118301423>. Citado 3 vezes nas páginas 9, 35, and 36.

PELLAND, S. et al. *Photovoltaic and Solar Forecasting: State of the Art*. [S.l.], 2013. Citado na página 24.

PERSSON, C. et al. Multi-site solar power forecasting using gradient boosted regression trees. *Solar Energy*, v. 150, p. 423 – 436, 2017. ISSN 0038-092X. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0038092X17303717>. Citado na página 43.

PODESTá, G. P. et al. Estimating daily solar radiation in the argentine pampas. *Agricultural and Forest Meteorology*, v. 123, n. 1, p. 41 – 53, 2004. ISSN 0168-1923. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0168192303002740>. Citado na página 44.

POLI, R.; LANGDON, W. B.; MCPHEE, N. F. *A field guide to genetic programming*. [S.l.]: Published via `http://lulu.com` and freely available at `http://www.gp-field-guide.org.uk`, 2008. (With contributions by J. R. Koza). Citado na página 67.

POLI, R.; MCPHEE, N. F.; VANNESCHI, L. Elitism reduces bloat in genetic programming. In: *Proceedings of the 10th annual conference on Genetic and evolutionary computation*. [S.l.: s.n.], 2008. Citado na página 67.

PROKOP, L. et al. Supervised learning of photovoltaic power plant output prediction models. *Neural Network World*, v. 23, n. 4, p. 321–338, 2013. Cited By 16. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-84885672403&doi=10.14311%2fNNW.2013.23.020&partnerID=40&md5=accbeab4c3475b4d2a7897458ef27590>. Citado na página 44.

RANA, M.; KOPRINSKA, I.; AGELIDIS, V. G. Univariate and multivariate methods for very short-term solar photovoltaic power forecasting. *Energy Conversion and Management*, v. 121, p. 380 – 390, 2016. ISSN 0196-8904. Citado 6 vezes nas páginas 12, 31, 32, 59, 65, and 72.

REIKARD, G. Predicting solar radiation at high resolutions: A comparison of time series forecasts. *Solar Energy*, v. 83, n. 3, p. 342 – 349, 2009. ISSN 0038-092X. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0038092X08002107>. Citado na página 44.

REIKARD, G.; HANSEN, C. Forecasting solar irradiance at short horizons: Frequency and time domain models. *Renewable Energy*, v. 135, p. 1270 – 1290, 2019. ISSN 0960-1481. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0960148118310310>. Citado 2 vezes nas páginas 39 and 74.

RENO, M. J.; HANSEN, C. W.; STEIN, J. S. *Global Horizontal Irradiance Clear Sky Models: Implementation and Analysis*. [S.l.], 2012. Citado na página 55.

RUSSO, M. et al. Genetic programming for photovoltaic plant output forecasting. *Solar Energy*, v. 105, p. 264 – 273, 2014. ISSN 0038-092X. Citado na página 61.

SAMANTA, M.; SRIKANTH, B. K.; YERRAPRAGADA, J. B. *Short-Term Power Forecasting of Solar PV Systems Using Machine Learning Techniques*. 2015. Citado na página 44.

SEARSON, D. P.; LEAHY, D. E.; WILLIS, M. J. Gptips: an open source genetic programming toolbox for multigene symbolic regression. In: CITESEER. *Proceedings of* *the International multiconference of engineers and computer scientists*. [S.l.], 2010. v. 1, p. 77–80. Citado na página 64.

SODA. *Solar Energy Services for Professionals*. 2017. Disponível em: <http://www.soda-pro.com/home;jsessionid=B032D33B0AD3460B741E14E41CC46BE2>. Citado na página 56.

SOLAR Lab UFG. Disponível em: <https://sites.google.com/site/sfvemcufg/>. Citado na página 49.

SOLAR Tech Lab. Disponível em: <http://www.solartech.polimi.it/>. Citado na página 50.

SPERATI, S. et al. The "weather intelligence for renewable energies" benchmarking exercise on short-term forecasting of wind and solar power generation. *Energies*, v. 8, n. 9, p. 9594–9619, 2015. ISSN 1996-1073. Disponível em: <https://www.mdpi.com/1996-1073/8/9/9594>. Citado na página 24.

TSO, G. K.; YAU, K. K. Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks. *Energy*, v. 32, n. 9, p. 1761 – 1768, 2007. ISSN 0360-5442. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0360544206003288>. Citado na página 44.

VOYANT, C. et al. Machine learning methods for solar radiation forecasting: A review. *Renewable Energy*, v. 105, p. 569 – 582, 2017. ISSN 0960-1481. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0960148116311648>. Citado 7 vezes nas páginas 12, 24, 28, 30, 39, 43, and 44.

WANG, N. et al. Life cycle carbon emission modelling of coal-fired power: Chinese case. *Energy*, v. 162, p. 841 – 852, 2018. ISSN 0360-5442. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0360544218315743>. Citado na página 23.

WOLFF, B.; LORENZ, E.; KRAMER, O. Statistical learning for short-term photovoltaic power predictions. In: *Computational sustainability*. [S.l.]: Springer, 2016. p. 31–45. Citado na página 44.

WU, J. et al. Prediction of solar radiation with genetic approach combing multi-model framework. *Renewable Energy*, v. 66, p. 132 – 139, 2014. ISSN 0960-1481. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0960148113006526>. Citado na página 44.

WU, Y.-K.; CHEN, C.-R.; RAHMAN, H. A. A novel hybrid model for short-term forecasting in pv power generation. *International Journal of Photoenergy*, Hindawi, v. 2014, 2014. Citado na página 44.

YAKOWITZ, S. Nearest-neighbour methods for time series analysis. *Journal of Time Series Analysis*, v. 8, n. 2, p. 235–247, 1987. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9892.1987.tb00435.x>. Citado na página 42.

YANG, D. et al. History and trends in solar irradiance and pv power forecasting: A preliminary assessment and review using text mining. *Solar Energy*, v. 168, p. 60 – 101, 2018. ISSN 0038-092X. Advances in Solar Resource Assessment and Forecasting. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0038092X17310022>. Citado 8 vezes nas páginas 24, 28, 30, 31, 33, 38, 41, and 45.

Yang, H. et al. A weather-based hybrid method for 1-day ahead hourly forecasting of pv power output. *IEEE Transactions on Sustainable Energy*, v. 5, n. 3, p. 917–926, July 2014. ISSN 1949-3037. Citado na página 44.

ZAMO, M. et al. A benchmark of statistical regression methods for short-term forecasting of photovoltaic electricity production, part i: Deterministic forecast of hourly production. *Solar Energy*, v. 105, p. 792 – 803, 2014. ISSN 0038-092X. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0038092X13005239>. Citado na página 44.

ZENDEHBOUDI, A.; BASEER, M.; SAIDUR, R. Application of support vector machine models for forecasting solar and wind energy resources: A review. *Journal of Cleaner Production*, v. 199, p. 272 – 285, 2018. ISSN 0959-6526. Disponível em: <http://www.sciencedirect.com/science/article/pii/S095965261832153X>. Citado na página 41.

Appendix

Table .1 – Forecast errors for Goiania (best values in bold). *Persist*: persistence.

FH	Method	**Haur. k_t^*** s	RMSE	σ_{RMSE}	MAE	σ_{MAE}	**Inei. k_t^*** s	RMSE	σ_{RMSE}	MAE	σ_{MAE}
	$Persist$		120.67		64.13			120.64		63.85	
	GP	14.51	103.16	0.28	60.32	**0.29**	14.29	103.40	**0.37**	**60.09**	0.62
15	ANN	**15.46**	**102.02**	**0.27**	**59.97**	0.44	**15.08**	**102.45**	**0.37**	60.19	**0.45**
	GP_{ens}	14.90	102.68		59.89		14.53	103.12		59.83	
	ANN_{ens}	**16.05**	**101.30**		**59.09**		**15.87**	**101.50**		**59.16**	
	$Persist$		151.11		85.42			150.81		84.59	
	GP	13.75	130.34	**0.52**	82.36	**0.61**	13.56	130.36	**0.29**	82.31	**0.37**
30	ANN	**14.87**	**128.64**	0.59	**80.49**	0.72	**14.49**	**128.95**	0.39	**81.02**	0.90
	GP_{ens}	14.11	129.79		81.84		13.94	129.78		81.81	
	ANN_{ens}	**15.52**	**127.65**		**79.34**		**15.28**	**127.76**		**79.54**	
	$Persist$		163.39		96.35			162.72		94.92	
	GP	15.46	138.13	**0.31**	90.12	**0.36**	14.93	138.43	**0.44**	90.94	**0.28**
45	ANN	**15.65**	**137.82**	0.37	**89.29**	0.40	**15.03**	**138.26**	0.50	**89.86**	0.65
	GP_{ens}	15.56	137.96		89.67		15.06	138.21		90.56	
	ANN_{ens}	**16.40**	**136.60**		**87.99**		**16.03**	**136.64**		**88.43**	
	$Persist$		170.94		103.64			169.82		101.44	
	GP	16.13	143.36	**0.40**	96.00	**0.42**	**15.92**	**142.78**	**0.26**	95.69	**0.46**
60	ANN	**16.21**	**143.22**	0.75	**93.98**	1.21	15.77	143.04	0.50	**94.52**	0.69
	GP_{ens}	16.45	142.82		95.61		16.07	142.53		95.44	
	ANN_{ens}	**17.03**	**141.83**		**92.46**		**16.36**	**142.04**		**93.90**	
	$Persist$		178.00		110.33			176.44		107.51	
	GP	**17.36**	**147.11**	**0.45**	99.50	**0.41**	**16.70**	**146.98**	**0.25**	99.56	**0.41**
75	ANN	17.34	147.15	0.66	**98.45**	0.79	16.42	147.47	0.52	**98.81**	0.50
	GP_{ens}	17.62	146.65		98.97		16.80	146.79		99.35	
	ANN_{ens}	**18.14**	**145.71**		**96.98**		**17.25**	**146.00**		**97.42**	
	$Persist$		185.30		117.29			183.28		113.91	
	GP	18.86	150.35	**0.40**	102.47	**0.45**	**18.18**	**149.96**	0.40	102.16	**0.63**
90	ANN	**19.08**	**149.95**	0.43	**101.11**	0.68	18.03	150.23	**0.28**	**101.49**	0.88
	GP_{ens}	19.09	149.91		102.13		18.35	149.65		101.73	
	ANN_{ens}	**19.88**	**148.46**		**99.57**		**18.84**	**148.76**		**99.91**	
	$Persist$		192.48		123.01			190.04		119.03	
	GP	20.71	152.61	**0.35**	104.21	**0.53**	**19.91**	**152.21**	**0.36**	104.13	**0.35**
105	ANN	**20.79**	**152.45**	0.50	**102.52**	0.56	19.45	153.09	0.81	**103.82**	1.30
	GP_{ens}	20.94	152.16		103.84		20.07	151.91		103.82	
	ANN_{ens}	**21.46**	**151.18**		**101.29**		**20.20**	**151.65**		**102.47**	
	$Persist$		199.62		128.73			196.80		124.08	
	GP	**22.65**	**154.40**	**0.27**	106.04	**0.32**	**21.58**	**154.33**	**0.37**	**105.85**	**0.49**
120	ANN	22.27	155.17	0.92	**105.71**	1.01	21.12	155.24	0.77	106.49	1.00
	GP_{ens}	22.80	154.10		105.79		**21.77**	**153.96**		105.50	
	ANN_{ens}	**23.09**	**153.52**		**104.23**		21.73	154.03		**105.26**	

Table .2 – Forecast errors for Milan (best values in bold).

		Haur. k_t^*					**Inei. k_t^***				
FH	**Method**	**s**	**RMSE**	$\boldsymbol{\sigma_{RMSE}}$	**MAE**	σ_{MAE}	s	**RMSE**	$\boldsymbol{\sigma_{RMSE}}$	**MAE**	$\boldsymbol{\sigma_{MAE}}$
	$Persist$		76.02		37.43			75.97		37.33	
	GP	11.63	67.18	**0.63**	**34.22**	0.38	11.35	67.34	**0.63**	**34.06**	0.30
15	ANN	**12.54**	**66.48**	1.09	35.11	**0.31**	**12.06**	**66.80**	1.04	36.11	**0.25**
	GP_{ens}	12.23	66.72		**33.84**		11.95	66.89		**33.69**	
	ANN_{ens}	**14.81**	**64.76**		34.11		**14.01**	**65.32**		34.18	
	$Persist$		99.41		51.12			99.29		50.80	
	GP	10.80	88.67	**0.32**	**48.95**	**0.44**	**9.62**	**89.74**	**0.35**	**50.18**	**0.38**
30	ANN	**10.86**	**88.61**	0.71	50.23	0.51	9.24	90.12	1.80	51.49	1.09
	GP_{ens}	11.12	88.36		**48.65**		10.12	89.24		49.73	
	ANN_{ens}	**12.62**	**86.87**		48.90		**11.82**	**87.55**		**48.92**	
	$Persist$		110.31		59.34			110.12		58.76	
	GP	**10.70**	**98.50**	**0.20**	**56.46**	**0.42**	**9.18**	**100.01**	**0.38**	**56.92**	**0.52**
45	ANN	10.40	98.83	0.64	58.27	0.64	8.51	100.75	0.81	59.04	0.65
	GP_{ens}	10.87	98.32		**56.25**		9.46	99.70		**56.63**	
	ANN_{ens}	**12.01**	**97.06**		56.94		**11.02**	**97.99**		57.17	
	$Persist$		120.03		66.08			119.82		65.28	
	GP	**11.58**	**106.12**	**0.22**	**62.04**	**0.49**	**11.09**	**106.52**	**0.35**	**62.78**	**0.39**
60	ANN	11.55	106.16	0.61	63.80	0.79	10.48	107.27	0.78	63.56	0.47
	GP_{ens}	11.83	105.83		**61.75**		11.52	106.02		62.70	
	ANN_{ens}	**12.87**	**104.58**		62.50		**11.91**	**105.54**		**61.69**	
	$Persist$		128.66		72.05			128.54		71.04	
	GP	**12.80**	**112.18**	**0.34**	**67.43**	**0.32**	**12.36**	**112.66**	**0.65**	**67.20**	**0.50**
75	ANN	12.67	112.36	0.90	67.97	0.72	11.76	113.42	1.48	69.28	1.15
	GP_{ens}	12.44	112.65		67.22		12.50	112.47		**67.03**	
	ANN_{ens}	**13.94**	**110.72**		**66.63**		**13.35**	**111.38**		67.63	
	$Persist$		136.10		77.66			136.15		76.32	
	GP	**13.60**	**117.59**	**0.40**	**70.97**	**0.29**	**13.93**	**117.18**	**0.66**	**71.16**	**0.49**
90	ANN	13.52	117.69	0.53	72.62	1.02	13.39	117.91	0.85	72.66	1.37
	GP_{ens}	13.87	117.22		**70.78**		14.35	116.61		70.86	
	ANN_{ens}	**14.79**	**115.97**		71.25		**15.08**	**115.62**		**70.74**	
	$Persist$		142.26		82.92			142.50		81.48	
	GP	**14.60**	**121.49**	**0.52**	**74.05**	**0.35**	**14.60**	**121.69**	**0.44**	**74.92**	**0.45**
105	ANN	13.64	122.85	2.41	76.34	1.21	14.16	122.33	0.89	76.22	0.70
	GP_{ens}	14.77	121.26		**73.85**		14.77	121.46		74.71	
	ANN_{ens}	**15.45**	**120.29**		74.41		**15.71**	**120.11**		**74.56**	
	$Persist$		147.60		87.74			148.21		86.14	
	GP	**15.70**	**124.43**	**0.36**	**76.90**	**0.28**	**15.95**	**124.58**	**0.36**	**78.46**	**1.07**
120	ANN	15.03	125.42	1.21	79.81	1.64	15.57	125.14	1.01	79.72	1.26
	GP_{ens}	15.83	124.24		**76.73**		16.13	124.30		78.28	
	ANN_{ens}	**16.25**	**123.61**		78.47		**16.97**	**123.06**		**78.15**	

Table .3 – Forecast errors for Desert Rock (best values in bold).

		Haur. k_t^*					**Inei. k_t^***				
FH	**Method**	**s**	**RMSE**	σ_{RMSE}	**MAE**	σ_{MAE}	**s**	**RMSE**	σ_{RMSE}	**MAE**	σ_{MAE}
	$Persist$		78.00		34.46			77.90		33.59	
	GP	11.96	68.68	**0.14**	**31.64**	**0.18**	11.94	68.60	**0.23**	**31.81**	**0.23**
15	ANN	**11.98**	**68.66**	0.25	31.85	0.56	**11.98**	**68.57**	0.29	31.84	0.51
	GP_{ens}	12.26	68.44		31.29		12.28	68.34		31.58	
	ANN_{ens}	13.24	67.68		31.02		12.96	67.81		31.27	
	$Persist$		101.32		47.74			101.05		45.88	
	GP	**11.45**	**89.72**	**0.57**	**45.24**	**0.41**	**11.80**	**89.12**	**0.09**	**44.82**	0.45
30	ANN	10.74	90.44	0.71	44.93	0.43	10.78	90.16	0.53	44.83	**0.37**
	GP_{ens}	11.77	89.39		44.73		11.91	89.01		44.68	
	ANN_{ens}	11.60	89.57		44.15		11.62	89.31		44.33	
	$Persist$		111.75		55.36			111.24		52.51	
	GP	**12.09**	**98.24**	**0.28**	**50.87**	**0.35**	**11.89**	**98.02**	1.41	**50.91**	1.08
45	ANN	10.98	99.49	0.46	50.94	1.00	10.95	99.06	**0.22**	51.10	**0.67**
	GP_{ens}	12.21	98.11		50.72		12.47	97.37		50.47	
	ANN_{ens}	11.85	98.51		50.00		11.53	98.41		50.51	
	$Persist$		118.32		61.63			117.51		57.80	
	GP	**13.05**	**102.88**	**0.32**	**55.24**	**0.38**	**13.13**	**102.09**	**0.15**	**54.46**	**0.16**
60	ANN	11.81	104.34	1.21	56.51	1.01	11.28	104.26	0.29	54.99	0.47
	GP_{ens}	13.17	102.74		55.11		13.19	102.02		54.41	
	ANN_{ens}	12.77	103.20		55.35		11.90	103.53		54.47	
	$Persist$		124.48		67.03			123.35		62.16	
	GP	**14.46**	**106.48**	**0.11**	**59.12**	**0.10**	**14.25**	**105.77**	**0.11**	**58.47**	**0.34**
75	ANN	13.34	107.88	0.33	59.70	1.50	12.74	107.63	0.47	58.80	0.60
	GP_{ens}	14.56	106.36		58.42		14.32	105.69		57.81	
	ANN_{ens}	14.06	106.98		58.31		13.41	106.81		57.42	
	$Persist$		129.33		71.66			127.87		65.93	
	GP	**15.18**	**109.70**	**0.20**	**61.55**	**0.36**	**14.77**	**108.98**	**0.15**	**60.57**	**0.24**
90	ANN	13.87	111.40	0.76	61.60	0.91	12.81	111.50	0.80	61.12	0.65
	GP_{ens}	15.25	109.60		61.47		14.84	108.90		60.44	
	ANN_{ens}	14.72	110.30		60.67		13.55	110.55		59.99	
	$Persist$		133.39		75.55			131.57		68.98	
	GP	**16.06**	**111.97**	**0.14**	**63.15**	**0.22**	**15.36**	**111.36**	**0.15**	**63.13**	**0.27**
105	ANN	14.22	114.42	0.70	63.20	0.89	13.68	113.57	0.39	63.42	0.60
	GP_{ens}	16.11	111.91		63.49		15.45	111.24		62.52	
	ANN_{ens}	14.99	113.40		62.11		14.34	112.70		62.02	
	$Persist$		137.62		79.56			135.42		72.19	
	GP	**16.83**	**114.46**	**0.20**	66.09	**0.25**	**15.85**	**113.96**	**0.15**	**65.14**	**0.29**
120	ANN	15.60	116.15	0.56	**65.59**	0.52	14.36	115.98	0.42	65.54	1.16
	GP_{ens}	16.91	114.35		65.98		15.91	113.88		65.07	
	ANN_{ens}	16.24	115.27		64.86		15.07	115.01		64.83	

Table .4 – Forecast errors for Pennsylvania State University (best values in bold).

		Haur. k_t^*					**Inei. k_t^***				
FH	**Method**	**s**	**RMSE**	σ_{RMSE}	**MAE**	σ_{MAE}	**s**	**RMSE**	σ_{RMSE}	**MAE**	σ_{MAE}
	Persist		94.43		51.76			94.37		51.49	
	GP	12.78	82.36	0.44	**47.47**	**0.34**	12.99	82.11	**0.28**	**47.23**	**0.14**
15	*ANN*	**14.42**	**80.82**	**0.29**	47.49	0.45	**13.86**	**81.29**	0.29	48.17	0.53
	GP_{ens}	13.31	81.86		47.10		13.39	81.73		**46.92**	
	ANN_{ens}	**14.98**	**80.28**		**47.00**		**14.49**	**80.69**		47.57	
	Persist		118.31		68.46			118.13		67.85	
	GP	10.71	105.64	0.26	66.56	**0.28**	10.46	105.78	**0.26**	**66.02**	**0.28**
30	*ANN*	**11.70**	**104.47**	**0.17**	**65.93**	0.61	**11.54**	**104.49**	0.36	66.47	0.37
	GP_{ens}	11.05	105.23		66.23		10.77	105.40		65.78	
	ANN_{ens}	**12.34**	**103.72**		**65.28**		**12.29**	**103.62**		**65.66**	
	Persist		131.38		78.48			131.05		77.51	
	GP	12.32	115.19	**0.23**	**74.60**	**0.40**	11.85	115.52	**0.21**	**74.30**	**0.29**
45	*ANN*	**13.07**	**114.21**	0.28	74.80	0.63	**12.71**	**114.39**	0.23	74.52	0.45
	GP_{ens}	12.61	114.82		74.26		12.06	115.24		74.01	
	ANN_{ens}	**13.76**	**113.30**		**74.07**		**13.44**	**113.44**		**73.69**	
	Persist		138.63		84.75			138.19		83.45	
	GP	12.54	121.25	**0.06**	79.96	**0.17**	12.17	121.37	**0.29**	80.29	**0.31**
60	*ANN*	**13.11**	**120.45**	0.36	**79.94**	0.87	**12.67**	**120.69**	0.41	**80.03**	0.72
	GP_{ens}	12.68	121.06		79.78		12.39	121.06		80.00	
	ANN_{ens}	**13.68**	**119.66**		**79.28**		**13.42**	**119.64**		**79.17**	
	Persist		144.19		90.43			143.68		88.84	
	GP	12.65	125.95	**0.11**	84.92	**0.21**	12.40	125.86	0.86	84.80	0.97
75	*ANN*	**12.88**	**125.62**	0.48	**84.87**	0.98	**12.57**	**125.63**	**0.42**	**84.71**	**0.40**
	GP_{ens}	12.86	125.65		84.71		12.77	125.33		84.45	
	ANN_{ens}	**13.57**	**124.63**		**84.10**		**13.27**	**124.61**		**83.81**	
	Persist		150.58		96.09			150.03		94.27	
	GP	**13.34**	**130.50**	**0.36**	**89.32**	**0.20**	**13.63**	**129.59**	**0.29**	**89.07**	**0.37**
90	*ANN*	13.27	130.59	0.61	89.37	0.98	13.06	130.43	0.50	89.50	0.68
	GP_{ens}	13.52	130.23		89.09		13.77	129.37		88.89	
	ANN_{ens}	**14.13**	**129.30**		**88.42**		**13.89**	**129.19**		**88.57**	
	Persist		156.99		101.08			156.42		99.04	
	GP	**14.97**	**133.48**	**0.13**	**92.24**	**0.21**	**14.69**	**133.44**	**0.38**	**92.10**	**0.43**
105	*ANN*	14.38	134.42	0.23	92.99	0.94	13.85	134.75	0.45	93.37	0.53
	GP_{ens}	**15.14**	**133.22**		**92.13**		**14.90**	**133.11**		**91.71**	
	ANN_{ens}	15.08	133.31		92.23		14.67	133.46		92.44	
	Persist		164.37		106.50			163.82		104.31	
	GP	**16.23**	**137.70**	**0.21**	**95.99**	**0.15**	**16.21**	**137.27**	**0.22**	**95.95**	**0.22**
120	*ANN*	15.50	138.90	0.82	97.39	0.93	15.08	139.12	0.74	97.70	1.02
	GP_{ens}	**16.36**	**137.48**		**95.85**		**16.37**	**137.00**		**95.73**	
	ANN_{ens}	16.25	137.66		96.57		15.91	137.75		96.73	

Table .5 – Forecast errors for Bondville (best values in bold).

FH	Method	**Haur. k_t^*** s	RMSE	σ_{RMSE}	MAE	σ_{MAE}	**Inei. k_t^*** s	RMSE	σ_{RMSE}	MAE	σ_{MAE}
	Persist		81.28		43.45			81.20		43.01	
	GP	10.43	72.81	0.77	41.51	**0.35**	10.85	72.39	**0.41**	**41.05**	**0.15**
15	*ANN*	**11.74**	**71.74**	**0.49**	**40.87**	0.59	**11.10**	**72.18**	**0.41**	41.26	0.34
	GP_{ens}	11.31	72.09		41.05		11.30	72.02		40.84	
	ANN_{ens}	**12.71**	**70.95**		**40.22**		**11.92**	**71.52**		**40.71**	
	Persist		101.49		57.67			101.25		56.79	
	GP	9.97	91.37	**0.47**	56.08	**0.56**	**10.23**	**90.89**	**0.28**	**56.38**	0.29
30	*ANN*	**10.16**	**91.18**	0.48	**55.92**	0.73	9.57	91.56	0.29	56.50	**0.21**
	GP_{ens}	10.35	90.98		55.75		**10.52**	**90.61**		56.15	
	ANN_{ens}	**10.94**	**90.38**		**55.32**		10.41	90.71		**55.78**	
	Persist		113.22		66.80			112.82		65.39	
	GP	9.83	102.09	0.52	64.38	**0.48**	**10.28**	**101.22**	**0.42**	**64.37**	**0.53**
45	*ANN*	**10.61**	**101.20**	**0.32**	**63.27**	0.68	9.82	101.74	0.89	64.49	1.34
	GP_{ens}	10.26	101.60		63.92		10.56	100.91		64.15	
	ANN_{ens}	**11.37**	**100.35**		**62.57**		**10.74**	**100.71**		**63.72**	
	Persist		121.55		73.53			120.99		71.70	
	GP	**11.65**	**107.39**	0.35	**69.37**	**0.42**	**11.16**	**107.48**	0.46	**69.76**	0.79
60	*ANN*	11.38	107.73	**0.33**	69.37	0.48	10.72	108.02	**0.37**	69.98	**0.54**
	GP_{ens}	12.06	106.90		68.94		**11.58**	**106.97**		69.30	
	ANN_{ens}	**12.15**	**106.79**		**68.64**		11.53	107.04		**69.15**	
	Persist		127.95		79.19			127.21		77.04	
	GP	**12.01**	**112.58**	**0.11**	**73.78**	**0.20**	**11.19**	**112.97**	**0.30**	**73.98**	**0.32**
75	*ANN*	11.52	113.21	0.49	74.30	0.83	10.62	113.71	1.01	74.83	1.01
	GP_{ens}	12.22	112.31		**73.47**		11.43	112.67		**73.70**	
	ANN_{ens}	**12.29**	**112.23**		73.56		**11.62**	**112.43**		73.77	
	Persist		134.49		84.73			133.56		82.22	
	GP	**13.22**	**116.71**	**0.19**	**77.69**	**0.21**	**12.15**	**117.33**	**0.22**	**78.67**	**0.28**
90	*ANN*	12.24	118.03	0.51	78.32	0.74	11.40	118.33	0.71	78.93	0.48
	GP_{ens}	**13.37**	**116.51**		**77.49**		12.29	117.14		78.49	
	ANN_{ens}	12.99	117.02		77.53		**12.44**	**116.94**		**77.77**	
	Persist		140.34		89.83			139.23		86.79	
	GP	**13.52**	**121.37**	**0.32**	**81.88**	**0.39**	**12.21**	**122.23**	**0.22**	**82.32**	**0.45**
105	*ANN*	12.56	122.71	0.69	82.06	0.52	11.90	122.66	0.55	82.70	0.45
	GP_{ens}	**13.87**	**120.87**		**81.01**		12.43	121.93		82.10	
	ANN_{ens}	13.40	121.53		81.18		**12.96**	**121.19**		**81.49**	
	Persist		145.88		95.01			144.58		91.41	
	GP	**14.42**	**124.84**	**0.35**	**85.95**	**0.31**	**12.96**	**125.85**	0.72	**86.30**	**0.55**
120	*ANN*	13.31	126.46	0.77	86.47	1.35	12.54	126.46	**0.55**	86.82	0.76
	GP_{ens}	**14.59**	**124.59**		**85.07**		13.25	125.42		85.79	
	ANN_{ens}	14.45	124.80		85.30		**13.59**	**124.93**		**85.66**	

Table .6 – Forecast errors for Sioux Falls (best values in bold).

		Haur. k_t^*					**Inei. k_t^***				
FH	**Method**	**s**	**RMSE**	σ_{RMSE}	**MAE**	σ_{MAE}	**s**	**RMSE**	σ_{RMSE}	**MAE**	σ_{MAE}
	$Persist$		75.47		40.51			75.40		40.06	
	GP	10.09	67.85	0.30	38.17	0.24	10.13	67.76	**0.23**	**37.82**	**0.19**
15	ANN	**12.25**	**66.23**	**0.20**	**38.02**	**0.18**	**11.70**	**66.58**	1.04	38.41	0.25
	GP_{ens}	10.46	67.58		37.91		10.32	67.61		**37.70**	
	ANN_{ens}	**13.04**	**65.63**		**37.55**		**12.76**	**65.78**		37.78	
	$Persist$		95.39		54.00			95.22		53.11	
	GP	9.22	86.60	**0.10**	**52.86**	**0.17**	8.71	86.93	0.28	52.97	**0.25**
30	ANN	**9.51**	**86.32**	0.45	53.15	0.51	**9.17**	**86.49**	**0.26**	**52.82**	0.42
	GP_{ens}	9.31	86.51		52.77		9.20	86.46		52.77	
	ANN_{ens}	**10.38**	**85.49**		**52.56**		**9.95**	**85.75**		**52.27**	
	$Persist$		107.46		62.87			107.17		61.44	
	GP	9.56	97.18	**0.21**	**60.86**	**0.42**	**9.72**	**96.75**	**0.09**	**60.69**	**0.42**
45	ANN	**9.84**	**96.89**	0.51	61.69	0.69	9.57	96.91	0.30	61.03	0.48
	GP_{ens}	9.79	96.94		**60.61**		10.00	96.45		**60.46**	
	ANN_{ens}	**10.62**	**96.05**		61.13		**10.23**	**96.20**		60.52	
	$Persist$		116.51		69.81			116.12		67.95	
	GP	**10.59**	**104.17**	**0.11**	**66.92**	**0.23**	**10.42**	**104.03**	**0.17**	67.09	**0.30**
60	ANN	10.16	104.68	0.38	67.47	0.33	10.20	104.27	0.45	**67.05**	0.46
	GP_{ens}	10.72	104.02		66.80		10.52	103.90		66.96	
	ANN_{ens}	**10.98**	**103.72**		**66.74**		**10.98**	**103.37**		**66.39**	
	$Persist$		123.64		75.45			123.21		73.35	
	GP	**11.18**	**109.82**	**0.20**	**71.76**	**0.33**	**11.01**	**109.64**	**0.14**	71.93	**0.39**
75	ANN	10.92	110.14	0.58	72.21	0.86	10.76	109.95	0.33	**71.58**	0.67
	GP_{ens}	11.37	109.58		71.56		11.18	109.43		71.73	
	ANN_{ens}	**11.75**	**109.11**		**71.49**		**11.51**	**109.03**		**70.92**	
	$Persist$		130.98		81.05			130.57		78.71	
	GP	**12.47**	**114.65**	**0.19**	**75.83**	**0.15**	**12.31**	**114.49**	**0.11**	**75.46**	**0.35**
90	ANN	11.84	115.48	0.42	76.45	1.00	11.60	115.42	0.27	76.37	0.76
	GP_{ens}	12.65	114.42		**75.62**		12.42	114.35		**75.33**	
	ANN_{ens}	**12.71**	**114.34**		75.66		**12.43**	**114.34**		75.59	
	$Persist$		138.49		86.65			138.10		83.89	
	GP	**13.92**	**119.20**	**0.15**	**79.62**	**0.34**	**13.70**	**119.18**	**0.25**	**79.31**	**0.21**
105	ANN	12.90	120.62	0.20	80.50	0.59	12.99	120.17	0.50	80.39	0.79
	GP_{ens}	**14.12**	**118.94**		**79.41**		**13.94**	**118.85**		**79.10**	
	ANN_{ens}	13.81	119.36		79.65		13.88	118.94		79.52	
	$Persist$		143.72		91.37			143.38		88.30	
	GP	**14.59**	**122.75**	**0.32**	**82.76**	**0.20**	**13.98**	**123.34**	0.92	**83.06**	1.03
120	ANN	13.71	124.02	0.63	84.05	0.48	13.34	124.26	**0.60**	83.99	**0.81**
	GP_{ens}	**14.76**	**122.50**		**82.52**		**14.34**	**122.82**		**82.70**	
	ANN_{ens}	14.52	122.85		83.29		14.23	122.97		83.11	