**INTER-FACULTY MASTER PROGRAM on
COMPLEX SYSTEMS and NETWORKS**

**SCHOOL of MATHEMATICS
SCHOOL of BIOLOGY
SCHOOL of GEOLOGY
SCHOOL of ECONOMICS**

ARISTOTLE UNIVERSITY of THESSALONIKI

Master Thesis

Title:

Forecasting power output of photovoltaic systems using machine learning techniques

Georgia Xanthopoulou

SUPERVISOR: Ioannis Antoniou, Professor, AUTH

CO-SUPERVISOR: Dionisios Kehagias, Researcher, CERTH/ITI

Thessaloniki, November 2017

**ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ
ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ στα
ΠΟΛΥΠΛΟΚΑ ΣΥΣΤΗΜΑΤΑ και ΔΙΚΤΥΑ**

**ΤΜΗΜΑ ΜΑΘΗΜΑΤΙΚΩΝ
ΤΜΗΜΑ ΒΙΟΛΟΓΙΑΣ
ΤΜΗΜΑ ΓΕΩΛΟΓΙΑΣ
ΤΜΗΜΑ ΟΙΚΟΝΟΜΙΚΩΝ ΕΠΙΣΤΗΜΩΝ**

**ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ
ΘΕΣΣΑΛΟΝΙΚΗΣ**

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Τίτλος Εργασίας:

Πρόβλεψη παραγόμενης ενέργειας σε φωτοβολταϊκά συστήματα με τη χρήση τεχνικών μηχανικής μάθησης

Γεωργία Ξανθοπούλου

ΕΠΙΒΛΕΠΩΝ: Ιωάννης Αντωνίου, Καθηγητής Α.Π.Θ.

ΣΥΝΕΠΙΒΛΕΠΩΝ: Διονύσιος Κεχαγιάς, Ερευνητής Βαθμίδας Γ’, ΕΚΕΤΑ/ΙΠΤΗΛ

Εγκρίθηκε από την Τριμελή Εξεταστική Επιτροπή την 27η Νοεμβρίου 2017.

....	....	....
Ι. Αντωνίου	Δ. Κεχαγιάς	Π. Μπαμίδης
Καθηγητής Α.Π.Θ.	Ερευνητής Βαθμίδας Γ’ ΕΚΕΤΑ/ΙΠΤΗΛ	Αν. Καθηγητής Α.Π.Θ.

Θεσσαλονίκη, Νοέμβριος 2017

…………………………………………..
Γεωργία Β. Ξανθοπούλου
Πτυχιούχος Μαθηματικός Πανεπιστημίου Ιωαννίνων

Οι απόψεις και τα συμπεράσματα που περιέχονται σε αυτό το έγγραφο εκφράζουν τον συγγραφέα και δεν πρέπει να ερμηνευτεί ότι εκφράζουν τις επίσημες θέσεις του Α.Π.Θ.

ABSTRACT

The subject of this master thesis is the development of machine learning techniques from meteorological data, in order to accurately predict the value of the power generated by a photovoltaic park. To achieve this goal, various techniques such as clustering and classification will be tested to detect current meteorological conditions and to derive through appropriate correlations the prediction of the generated energy at a specific time. Our development consists of the following steps: data pre-processing, application of advanced techniques, and finally evaluation to identify those parameters that affect the quality of the forecast. Accurate power output forecasting is a critical credibility factor for both conventional and renewable modern power systems. Renewable power systems, like photovoltaic (PV) systems, could be severely affected by alternating weather conditions, which have an important impact the forecast accuracy. In this thesis a comparative analysis between contemporary linear and non-linear methods for power output forecasting is provided. In particular, the Autoregressive Integrated Moving Average (ARIMA) model is used as a linear method and an Artificial Neural Network (ANN) as a non-linear one. Moreover, enhanced models that incorporate, apart from energy, meteorological variables as explanatory variables in both linear and non-linear models are presented. Preliminary results, through experimentation on real data from a photovoltaic park in Crete, Greece have shown that the proposed enhanced methods result in increased forecasting accuracy to the base models.

KEY WORDS

Autoregressive Integrated Moving Average**,** Artificial Neural Network, Time Series Forecasting**,** Photovoltaic Systems, Power Output Forecasting, Comparative Analysis

ΠΕΡΙΛΗΨΗ

Αντικείμενο της παρούσας διπλωματικής εργασίας είναι η ανάπτυξη τεχνικών μηχανικής μάθησης από μετεωρολογικά δεδομένα, με σκοπό την ακριβή πρόβλεψη της ενέργειας που παράγει ένα φωτοβολταϊκό πάρκο. Για την επίτευξη του στόχου αυτού στα πλαίσια αυτής της διπλωματικής εργασίας έχουν δοκιμαστεί διάφορες τεχνικές όπως αυτές της μηχανικής μάθησης, αλλά και τεχνικές μοντελοποίησης χρονοσειρών. Γίνεται ανίχνευση των τρεχουσών μετεωρολογικών συνθηκών και από αυτές να προκύψει, μέσω κατάλληλων συσχετίσεων, η πρόβλεψη της παραγόμενης ενέργειας σε συγκεκριμένο χρονικό ορίζοντα. Η ανάπτυξη περιλαμβάνει τη φάση της προ-επεξεργασίας των δεδομένων, της εφαρμογής των ανεπτυγμένων τεχνικών αλλά και της δοκιμαστικής σύγκρισης με σκοπό τον εντοπισμό των σημαντικών παραμέτρων που επηρεάζουν την ποιότητα της πρόβλεψης. Η ακριβής πρόβλεψη της εξερχόμενης ενέργειας είναι ένας κρίσιμος συντελεστής αξιοπιστίας τόσο για συμβατικά όσο και για ανανεώσιμα σύγχρονα συστήματα ενέργειας. Τα συστήματα ανανεώσιμων πηγών ενέργειας, όπως τα φωτοβολταϊκά συστήματα, επηρεάζονται σοβαρά από τις εναλλασσόμενες καιρικές συνθήκες, και αυτό είναι ένα ζήτημα που επηρεάζει την ακρίβεια των προβλέψεων. Στην παρούσα διπλωματική εργασία παρέχεται μια συγκριτική ανάλυση μεταξύ σύγχρονων γραμμικών και μη γραμμικών μεθόδων για την πρόβλεψη της ενέργειας. Συγκεκριμένα, το αυτοπαλινδρομούμενο μοντέλο κινούμενου μέσου χρησιμοποιείται ως γραμμική μέθοδος και ένα τεχνητό νευρωνικό δίκτυο ως ένα μη γραμμικό. Επιπλέον, παρουσιάζονται βελτιωμένα μοντέλα που ενσωματώνουν, εκτός από την ενέργεια, τις μετεωρολογικές μεταβλητές ως ανεξάρτητες μεταβλητές τόσο στην περίπτωση των γραμμικών όσο και στην περίπτωση των μη γραμμικών μοντέλων. Τα προκαταρκτικά αποτελέσματα, μέσω πειραμάτων που διεξήχθησαν πάνω σε πραγματικά δεδομένα από ένα φωτοβολταϊκό πάρκο στην Κρήτη, δείχνουν αυξημένη ακρίβεια πρόβλεψης των βελτιωμένων μεθόδων σε σύγκριση με τα βασικά μοντέλα.

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ

Αυτοπαλιδρομικό Μοντέλο Κινούμενου Μέσου, Τεχνητό Νευρωνικό Δίκτυο, Πρόβλεψη Χρονοσειρών, Φωτοβολταϊκό Πάρκο, Πρόβλεψη Παραγόμενης Ενέργειας, Συγκριτική Ανάλυση

CONTENTS

ΣΥΝΟΨΗ

Καθώς ο πληθυσμός του πλανήτη αυξάνεται καθημερινά, αυξάνονται και οι ανάγκες για παραγωγή ενέργειας. Οι ανανεώσιμες πηγές ενέργειας είναι ίσως το κλειδί για την πραγματοποίηση αυτής της ανάγκης. Η ηλιακή ενέργεια, που είναι άφθονη σε πολλά σημεία της Γης και πόσο μάλλον στη χώρα μας, αποτελεί μία σημαντική πηγή ενέργειας προς εκμετάλλευση. Η πρόβλεψη παραγόμενης ενέργειας από φωτοβολταϊκά πάρκα αποτελεί μέρος πολλών ερευνών και για την επίτευξή της χρησιμοποιούνται διάφορες τεχνικές όπως οι γραμμικές μέθοδοι, οι μη-γραμμικές και οι υβριδικές. Σε ορισμένες από αυτές γίνεται λόγος για τη σχέση των μετεωρολογικών συνθηκών και της ενέργειας ως παραγόμενο προϊόν, χωρίς όμως να χρησιμοποιούνται μετεωρολογικά και περιβαλλοντικά δεδομένα ως επιπρόσθετα στοιχεία εισόδου στα μοντέλα πρόβλεψης. Η συνηθέστερη τακτική είναι αυτή της ομαδοποίησης των μετεωρολογικών συνθηκών και η δημιουργία διαφορετικών μοντέλων για κάθε περίπτωση. Παραδείγματος χάριν διαφορετικό μοντέλο για τις βροχερές μέρες και διαφορετικό για τις ηλιόλουστες.

Η παρούσα μεταπτυχιακή εργασία στοχεύει στην παρουσίαση και κατ’ επέκταση σύγκριση μοντέλων που όχι μόνο προβλέπουν την παραγόμενη ενέργεια χρησιμοποιώντας ενεργειακά δεδομένα, αλλά εκμεταλλεύονται μετεωρολογικά και περιβαλλοντικά δεδομένα. Αρχικά επεξεργαστήκαμε το αρχικό σύνολο των δεδομένων γεμίζοντας τις τιμές που πιθανόν να έλειπαν και αφαιρώντας τις ακραίες τιμές. Έχοντας πλέον έτοιμα τα δεδομένα συνεχίσαμε στη δημιουργία χρονοσειρών που θα χρησιμοποιούσαμε για να πραγματοποιήσουμε τις προβλέψεις. Οι προβλέψεις μας ανήκουν σε δύο κατηγορίες: στη βραχυπρόθεσμη πρόβλεψη και τη μακροπρόθεσμη πρόβλεψη. Για να προβλέψουμε την παραγόμενη ενέργεια χρησιμοποιήσαμε και γραμμικά και μη γραμμικά μοντέλα και ως ανεξάρτητες μεταβλητές ορίσαμε τιμές ενέργειας. Επίσης, όπως προαναφέραμε, δημιουργήσαμε και μοντέλα που ως ανεξάρτητες μεταβλητές είχαν, εκτός από ενεργειακά, και μετεωρολογικά δεδομένα. Έπειτα συγκρίνοντας τα αποτελέσματα με βάση την ακρίβεια της πρόβλεψης καταλήξαμε στο συμπέρασμα πως τα μη γραμμικά μοντέλα οδηγούν σε καλύτερα αποτελέσματα σε σχέση με τα γραμμικά μοντέλα και στις δυο περιπτώσεις πρόβλεψης.

Τα διαθέσιμα δεδομένα για τη δημιουργία πρόβλεψης αντιστοιχούν σε 380 ημέρες. Συγκεκριμένα, αναφέρονται στο διάστημα από 1 Ιουλίου του 2013 μέχρι και τις 16

Ιουλίου του 2014. Για την κάθε ημέρα διατίθενται μετεωρολογικά δεδομένα όπως η ένταση της ηλιακής ακτινοβολίας και δεδομένα μετρήσεων σχετικά με χαρακτηριστικά των πάνελ, όπως η ισχύς του πάνελ. Οι μετρήσεις αφορούν φωτοβολταϊκό πάρκο που βρίσκεται στην περιοχή της Κρήτης και για κάθε ημέρα διατίθενται δεδομένα για 6 πάνελ του ίδιου τύπου.

Στο πρώτο μέρος της διπλωματικής εργασίας παρουσιάζεται η ανάλυση δεδομένων για 7 μέρες με σκοπό τη διερεύνηση πιθανής επιρροής της πρόβλεψης εξαιτίας χρήσης δεδομένων από διαφορετικά πάνελ. Σε κάθε χρονικό διάστημα διάρκειας 15 λεπτών κατά τη διάρκεια μιας ημέρας, αντιστοιχεί μία τιμή παραγόμενης ενέργειας και συνεπώς για κάθε μέρα έχουμε 96 τιμές. Χρησιμοποιούμε τιμές από την αυγή μέχρι τη δύση του ηλίου και κατά συνέπεια δεν χρησιμοποιούνται στην ανάλυση και τα 96 15-λεπτα, αλλά κατά μέσο όρο 55 από αυτά. Έπειτα από έλεγχο των δεδομένων μας για έλλειψη τιμών και πιθανή ύπαρξη ακραίων τιμών και στη συνέχεια αφαίρεση αυτών, δημιουργήσαμε τις χρονοσειρές που χρειαζόμαστε για την πρόβλεψη. Ακολούθησε η επεξεργασία αυτών των χρονοσειρών για περαιτέρω ανάλυση. Στη συνέχεια, γίνεται πρόβλεψη με χρήση αυτοπαλινδρομικών μοντέλων κινούμενου μέσου. Για αυτό το σκοπό πραγματοποιήθηκε αρχικά έλεγχος στασιμότητας των χρονοσειρών. Όσες από αυτές κρίθηκαν μη στάσιμες, σύμφωνα με τα αποτελέσματα του στατιστικού τεστ Augmented Dicker Fuller test, μετατράπηκαν σε στάσιμες χρησιμοποιώντας πρώτες διαφορές. Εκτός από το στατιστικό τεστ για τον έλεγχο της στασιμότητας υπολογίσαμε συνοπτικά στατιστικά στοιχεία και πραγματοποιήσαμε και οπτικό έλεγχο των γραφημάτων για διασταύρωση των αποτελεσμάτων. Συνεχίσαμε την ανάλυση με τη εύρεση της κατάλληλης τάξης για το αυτοπαλινδρομούμενο μοντέλο κινούμενου μέσου. Τα γραφήματα της συνάρτησης αυτοσυσχέτισης και της μερικής συνάρτησης αυτοσυσχέτισης μας βοήθησαν να επιλέξουμε την τάξη του αυτοπαλινδρομικού μοντέλου κινούμενου μέσου. Λόγω των γραφημάτων οδηγηθήκαμε σε χρήση τάξης 1, όμως για πιο ενδελεχή έρευνα αναπτύξαμε και μοντέλα τάξεως 2 και 3.

Στο επόμενο βήμα, δημιουργήσαμε τεχνητό νευρωνικό δίκτυο για την πρόβλεψη της παραγόμενης ενέργειας. Η τοπολογία του νευρωνικού δικτύου που χρησιμοποιήσαμε είναι: 1, 2 ή 3 νευρώνες για την είσοδο (αντίστοιχα με τις τάξεις του γραμμικού μοντέλου), ένα κρυφό στρώμα με 3 νευρώνες και ένας νευρώνας για την έξοδο. Για την επιλογή του αριθμού των νευρώνων στο κρυφό στρώμα διενεργήθηκε μια σειρά προσομοιώσεων κατά τις οποίες δημιουργήθηκαν νευρωνικά δίκτυα με διαφορετικό αριθμό νευρώνων στο κρυφό στρώμα, και πραγματοποιήθηκαν προβλέψεις για κάθε ένα από αυτά. Επιλέχθηκε τελικά ο αριθμός νευρώνων κρυφού στρώματος εκείνου του δικτύου που παρουσίασε τα καλύτερα αποτελέσματα πρόβλεψης. Πραγματοποιήσαμε και ανάλογη διερεύνηση σχετικά με τις συναρτήσεις ενεργοποίησης που χρησιμοποιούνται στο κρυφό στρώμα και καταλήξαμε ότι η σιγμοειδής συνάρτηση ενεργοποίησης δίνει τα καλύτερα αποτελέσματα. Επίσης χρησιμοποιήσαμε το στατιστικό τεστ Wilcoxon για να επιβεβαιώσουμε πως η επιλογή των νευρώνων κατά την είσοδο δεν ήταν τυχαία.

Όπως προαναφέραμε, εκτός από τις προβλέψεις για την παραγόμενη ενέργεια με χρήση μόνο ενεργειακών δεδομένων, έγινε και χρήση μετεωρολογικών και περιβαλλοντικών δεδομένων ως εισόδων στα μοντέλα πρόβλεψης. Με τη συμβολή αυτών, δημιουργήσαμε βελτιωμένα μοντέλα με σκοπό τη διερεύνηση της βελτίωσης ή μη των προβλέψεων σε κάθε περίπτωση. Συγκεκριμένα, χρησιμοποιήθηκαν τα ακόλουθα δεδομένα για τη δημιουργία των βελτιωμένων μοντέλων: η ηλιακή ακτινοβολία, η θερμοκρασία της ατμόσφαιρας και η θερμοκρασία του φωτοβολταϊκού πάνελ. Η επιλογή αυτών των τριών επιπλέον κατηγοριών δεδομένων έγινε βάσει των συσχετίσεων Pearson και Spearman μεταξύ αυτών και της μεταβλητής εξόδου, της ενέργειας. Πράγματι, οι τιμές των συσχετίσεων που προέκυψαν δείχνουν ισχυρή συσχέτιση, γεγονός που δικαιολογεί την επιλογή τους. Όπως και στα βασικά μοντέλα, έτσι και στα βελτιωμένα, χρησιμοποιήθηκαν γραμμικά και μη γραμμικά μοντέλα της ίδιας τάξης για την εξαγωγή των αποτελεσμάτων.

Για την πραγματοποίηση της πρόβλεψης έπρεπε να χωρίσουμε τα δεδομένα μας σε δύο σύνολα, το σύνολο εκπαίδευσης και το σύνολο ελέγχου, όπου το πρώτο χρησιμοποιήθηκε για την εκμάθηση του μοντέλου και το δεύτερο για την αξιολόγηση του μοντέλου. Το σύνολο εκπαίδευσης αποτελείται από το 80% των τιμών του αρχικού συνόλου δεδομένων και το σύνολο ελέγχου από το 20%. Οι προβλέψεις έγιναν με βήμα το διάστημα 15 λεπτών με χρονικό ορίζοντα 1 ώρας. Η διαδικασία επαναλήφθηκε και για τα 6 φωτοβολταϊκά πάνελ ώστε να επιτευχθεί διασταυρούμενη επικύρωση των αποτελεσμάτων. Τέλος, χρησιμοποιήσαμε το συμμετρικό μέσο σφάλμα απόλυτου ποσοστού (SMAPE) για τη σύγκριση των αποτελεσμάτων των βασικών και των βελτιωμένων μοντέλων.

Τα αποτελέσματα που συλλέξαμε από τις προβλέψεις που πραγματοποιήσαμε μας οδήγησαν στο συμπέρασμα πως το αυτοπαλινδρομούμενο μοντέλο τάξεως 1 δίνει καλύτερες προβλέψεις σε σύγκριση με τα μοντέλα τάξεως 2 και 3. Αυτό το γεγονός επιβεβαιώνει την ένδειξη χρήσης αυτοπαλινδρομικού μοντέλου τάξης 1, σύμφωνα με τα διαγράμματα των αυτοσυσχετίσεων. Όσον αφορά τα νευρωνικά δίκτυα, και εκεί τα δίκτυα με 1 νευρώνα για είσοδο δίνουν τα καλύτερα αποτελέσματα συγκριτικά με τους ανταγωνιστές τους. Αναφορικά με τη σύγκριση των δύο μοντέλων, αυτοπαλινδρομικού και νευρωνικού, το δεύτερο παρουσιάζει το μικρότερο σφάλμα και κατ' επέκταση δίνει καλύτερα αποτελέσματα πρόβλεψης. Συνεχίζοντας την ανάλυσή μας και με τα βελτιωμένα μοντέλα, παρατηρούμε πως η είσοδος επιπλέον μεταβλητής μειώνει την ακρίβεια των γραμμικών μοντέλων, ενώ αυξάνει την ακρίβεια των μη γραμμικών.

Με την προαναφερθείσα ανάλυση ολοκληρώνεται το πρώτο μέρος της εργασίας που ουσιαστικά είναι μια βραχυπρόθεσμη πρόβλεψη απόδοσης ενός φωτοβολταϊκού πάρκου. Το δεύτερο κομμάτι της εργασίας ασχολείται με την μακροπρόθεσμη πρόβλεψη της μέσης ημερήσιας απόδοσης ενός φωτοβολταϊκού πάρκου. Από τα αρχικά δεδομένα παρήχθη, με διαφορετική επεξεργασία, η χρονοσειρά με την οποία πραγματοποιήθηκαν οι προβλέψεις. Πραγματοποιήσαμε ξανά συμπλήρωση ελλειπόντων τιμών και αφαίρεση των ακραίων τιμών όπως και στη βραχυπρόθεσμη πρόβλεψη. Έπειτα, για κάθε ημέρα υπολογίσαμε τη μέση τιμή απόδοσης των 6 πάνελ, και έτσι δημιουργήσαμε τη χρονοσειρά μας. Χρησιμοποιώντας ένα φίλτρο Butterworth τάξεως 3, χωρίσαμε τη χρονοσειρά σε δύο περιόδους. Η πρώτη περίοδος περιλαμβάνει τιμές από την 1η Ιουλίου 2013 μέχρι την 16η Ιουλίου 2014, χωρίς τις τιμές από την 4η Νοεμβρίου 2013 μέχρι και την 13η Μαρτίου 2014 που απαρτίζουν την δεύτερη περίοδο. Την πρώτη περίοδο την ονομάσαμε «*καλοκαίρι*» και τη δεύτερη «*χειμώνα*». Η πρόβλεψη έγινε για κάθε περίοδο χωριστά και όπως και στο πρώτο μέρος της εργασίας, έτσι και εδώ αναπτύξαμε γραμμικά, μη γραμμικά και βελτιωμένα μοντέλα. Στηριχτήκαμε πάνω στην ίδια λογική για τη δημιουργία όλων των μοντέλων.

Σχετικά με το γραμμικό μοντέλο ARIMA ακολουθήσαμε τις ίδιες μεθόδους με την παραπάνω ανάλυση και καταλήξαμε στο συμπέρασμα ότι θα μπορούσαμε να χρησιμοποιήσουμε είτε τάξη 1, είτε τάξη 2, είτε τάξη 3 και επιλέγαμε τη βέλτιστη ανάλογα με τα αποτελέσματα. Πραγματοποιήθηκε ξανά έλεγχος στασιμότητας που μας έδειξε ότι η χρονοσειρά μας ήταν ήδη στάσιμη και έτσι μπορούσαμε να συνεχίσουμε στη δημιουργία των μοντέλων χωρίς περαιτέρω επεξεργασία των δεδομένων μας.

Εφόσον έπρεπε να ασχοληθούμε με αυτές τις 3 τάξεις, κρατήσαμε το ίδιο σκεπτικό και στη δημιουργία των νευρωνικών μας δικτύων με σκοπό την επερχόμενη συγκριτική τους μελέτη. Έπειτα επαναλάβαμε το στατιστικό τεστ Wilcoxon για τα καινούργια νευρωνικά δίκτυα ώστε να επιβεβαιώσουμε την μη τυχαία επιλογή νευρώνων κατά την είσοδο.

Φυσικά, από αυτήν την ανάλυση δεν θα έλειπε και η δημιουργία μοντέλων πρόβλεψης με την προσθήκη μετεωρολογικών και περιβαλλοντικών μεταβλητών. Όπως και στη βραχυπρόθεσμη ανάλυση, έτσι και εδώ δημιουργήθηκαν τρία βελτιωμένα μοντέλα που εκτός από ενεργειακές τιμές περιλαμβάνουν και τιμές ηλιακής ακτινοβολίας, θερμοκρασίας περιβάλλοντος και θερμοκρασίας του πάνελ.

Στο τμήμα της πρόβλεψης χωρίσαμε τα δεδομένα μας επίσης κατά ένα σύνολο που αποτελεί το 80% του συνόλου των δεδομένων για την εκπαίδευση του μοντέλου και κατά ένα σύνολο που αποτελείται από το υπόλοιπο 20% για την αξιολόγηση του, ενώ επιλέχθηκε η ίδια μετρική SMAPE για την ερμηνεία των αποτελεσμάτων μας. Στα μοντέλα προβλέψεων που χρησιμοποιήσαμε μόνο ενεργειακά δεδομένα, παρατηρούμε μία μικρή υπεροχή των νευρωνικών δικτύων, όπως και στη βραχυπρόθεσμη ανάλυση. Το κατάλληλο γραμμικό μοντέλο για τις προβλέψεις μας είναι το αυτοπαλινδρομούμενο τάξης 2, τόσο για την περίοδο «καλοκαίρι», όσο και για την περίοδο «χειμώνα». Όσον αφορά τα νευρωνικά δίκτυα η καταλληλότερη τοπολογία είναι αυτή με 1 νευρώνα κατά την είσοδο, 20 νευρώνες στο κρυφό στρώμα και 1 κατά την έξοδο. Για την περίοδο «καλοκαίρι» χρησιμοποιούμε τη συνάρτηση υπερβολικής εφαπτομένης ως συνάρτηση ενεργοποίησης και για την περίοδο «χειμώνας» τη σιγμοειδή συνάρτηση. Η επιλογή των συναρτήσεων ενεργοποίησης έγινε μετά από διερεύνηση, όπως και στο πρώτο μέρος της εργασίας.

Η πρόβλεψη των βελτιωμένων μοντέλων διαφοροποιείται ως προς τις συναρτήσεις ενεργοποίησης του κρυφού στρώματος του νευρωνικού δικτύου, ενώ στα γραμμικά μοντέλα παρατηρούμε πως και εδώ το μοντέλο δεύτερης τάξης δίνει τα καλύτερα αποτελέσματα. Κάθε βελτιωμένο μοντέλο, είναι πιο ακριβές με διαφορετική συνάρτηση ενεργοποίησης και διαφορετικό αριθμό νευρώνων στο κρυφό στρώμα. Συγκεκριμένα, το μοντέλο ενέργειας-ηλιακής ακτινοβολίας παρουσιάζει μικρότερο σφάλμα έχοντας για συνάρτηση ενεργοποίησης την συνάρτηση ανορθωμένης γραμμικής μονάδας και 20 νευρώνες στο κρυφό στρώμα για την περίοδο «καλοκαίρι». Το μοντέλο ενέργειας – θερμοκρασίας δίνει ακριβέστερα αποτελέσματα όταν η συνάρτηση ενεργοποίησης είναι η συνάρτηση ανορθωμένης γραμμικής μονάδας, όμως με 30 νευρώνες στο κρυφό στρώμα, ενώ το μοντέλο ενέργειας – θερμοκρασίας πάνελ όταν η συνάρτηση ενεργοποίησης είναι η σιγμοειδής και με 90 νευρώνες στο ενδιάμεσο στρώμα. Στην περίοδο «χειμώνα» το μοντέλο ενέργειας – ηλιακής ακτινοβολίας υπερέχει των ανταγωνιστών του με 20 νευρώνες στο κρυφό στρώμα και συνάρτηση ενεργοποίησης τη σιγμοειδή, ενώ το μοντέλο ενέργειας – θερμοκρασίας με 50 νευρώνες στο κρυφό στρώμα και συνάρτηση ενεργοποίησης την συνάρτηση υπερβολικής εφαπτομένης. Τέλος, το μοντέλο ενέργειας – θερμοκρασίας πάνελ παρέχει μεγαλύτερη ακρίβεια με τη σιγμοειδή συνάρτηση και 90 νευρώνες.

Παρότι οι συναρτήσεις ενεργοποίησης ποικίλουν, παρατηρούμε πως η συμβολή περιβαλλοντικών μεταβλητών για την πρόβλεψη παραγόμενης ισχύος ενισχύει τα νευρωνικά δίκτυα και δίνει αποτελέσματα με μικρότερο σφάλμα, ενώ στην περίπτωση των γραμμικών μοντέλων πρόβλεψης παρατηρείται μια μικρή μείωση της ακρίβειας. Γεγονός που ισχύει και στις δύο διαφορετικές περιόδους που αναλύουμε. Η ακρίβεια των νευρωνικών και των γραμμικών μοντέλων της περιόδου «χειμώνας» είναι μικρότερη σε σύγκριση με τα αποτελέσματα της περιόδου «καλοκαίρι», διότι το πλήθος των δεδομένων της περιόδου «χειμώνας» είναι αρκετά μικρότερο από την περίοδο «καλοκαίρι».

INTRODUCTION

Nowadays, while the number of global population is increasing, the energy consumption is also growing. The renewable energy sources are being used as power producers for many decades now and their contribution to energy economy is an indisputable fact. For the purpose of fulfilling the increased power demand, renewable power systems are globally deployed, reducing the negative environmental footprint of the conventional power systems. Supplementally, this contributes to reducing CO2 emission gases. Power sources as the solar irradiance, can produce electric power directly from sunlight without fuel consumption, which is the main reason that photovoltaic systems (PV) consist one of the most important devices in the field of exploitation of the renewable energy sources. Solar energy resources are abundant in several places on earth and as a result the installation of photovoltaic panels have been increased. Among energy sources, solar energy has the greatest energy potential. Nevertheless, solar radiation is affected by weather conditions and as a fact the power produced of photovoltaic systems depends on that. Several methods have been proposed in the literature regarding the task of power output forecasting in PV systems, not taking, however, into account the effect of the alternating weather conditions. Models that take into account the weather conditions may present interesting results.

Many researchers focus on providing a forecasting tool in order to predict PV power output with good accuracy. Forecasts are the key to reliable power system stabilization and output estimation. It has been referred that photovoltaic power prediction consists an important way to guarantee the stability for grid-connected photovoltaic power generation [[18]]. Moreover, it is useful for energy storage management and maintenance [[17]]. Those who work on time series forecasting and almost everyone in the forecasting literature agree that no single method is best in every situation, due to the fact that most of the problems that we are dealing to forecast are complex problems. Over the years several methods to time series forecasting have been proposed. These methods can be roughly classified into three major categories: linear, non-linear and hybrid. Linear methods, such as the Autoregressive Integrated Moving Average (ARIMA) models, the Autoregressive Moving Average (ARMA) models, the simple linear regression and the multiple linear regression models are traditionally used and are the most popular methods due to their statistical properties. From the opposing point of view, we refer to non-linear methods, we allude to Artificial Neural Networks (ANNs), Support Vector Machines (SVM), k Nearest Neighbour (kNN) etc. Also, noticeable research activity has focused on the development of hybrid methods, aiming at consolidating the favourable attributes of both linear and non-linear methods.

Accurate power output forecasting could increase the reliability and performance of the PV systems, and also prevent unnecessary operating costs. Therefore, the implementation of accurate, either linear or non-linear models, is of paramount importance. This task was implemented with linear models by several researchers such as Hamid Oudjana et al. [[5]]. In their study regression is used as a forecasting method because of short execution time contrary to neural networks. Also, this method requires a mathematical model, instead of neural networks that don't require one. Jiahao et al. [[6]] used linear regression to study power output characteristics. Accurate prediction with multiple linear regression is the aim of Oussama et al. [[1]] too. Hugo T.C Pedro et al. [[14]] who also used ARIMA and Persistent models as forecasting tools in order to achieve solar power output predictions. They came to the conclusion that no exogenous data such as solar irradiance telemetry is needed for forecasting. Solar panels themselves are efficient to be used for forecasting. Multiregression analysis is implemented by Maria Grazia De Giorgi et al.[[12]] to obtain a relationship between PV power and weather parameters.

Many researchers working on power output forecasting utilize models based on neural networks because of the non-linearity of the meteorological data (Oussama and Farah [[1]], Hamid et al. [[5]] and Mellit et al. [[13]]). In particular Mellit et al. [[13]] used two different ANNs depending on the classification of the days as cloudy or sunny, while Jie Shi et al.[[16]] based their predictions categorizing the weather conditions into 4 clusters. Preliminary weather classification was used by Chen et al.[[3]] too and indicated improvement in the forecasting accuracy. Nevertheless, none of the above have implemented new, improved models that include meteorological data as explanatory variables. Jiahao et al. [[6]] and Valerio Lo Brano et al. [[10]] investigated ANNs with different topologies in order to make more accurate forecasts, whereas Anil Rai et al. [[15]] analyzed ANNs accuracy by comparing different test datasets. Leva et al. [[9]] analyzed the sensitivity of the ANNs in power forecasting and Teo et al. [[17]] apart from the accuracy and the sensitivity of the ANNs, analyzed also the efficiency of different activation functions. Other studies provide comparative results of different models such as ANNs, ARIMA, SVM etc. [[1] , [5] , [16]]. As mentioned before, Hugo et al.[[14]] used linear models while conducted ANNs and kNN methods to compare them with the linear methods. In their study the methods that belong in the artificial intelligence family of methods performed better than other techniques. Hybrid models are also proposed by several authors indicating increased robustness. [[4] , [20]].

The main purpose of this study is to forecast the power output of photovoltaic systems through a comparative analysis between linear and non-linear models and to produce new and improved models to achieve better accuracy. In particular, the Autoregressive Integrated Moving Average (ARIMA) family of models from time series analysis are the linear evaluated models, and the Artificial Neural Networks (ANN) with different topologies (i.e. number of neurons at the hidden layer, different activation functions, etc.) are the non-linear. Both types of models use the energy itself as independent (input) variable. Additional, a set of both linear and non-linear models that use as independent variable not only the energy but also the meteorological variables (e.g. irradiance, panel temperature, etc.) are implemented and evaluated. These models consist the improved models that we are presenting in this thesis.

The rest of this thesis is organized as follows. Firstly, the thesis is divided into two parts. In the first part, there is an analysis of the 6 PV panels in a period of 7 days, which represent a short term forecasting analysis and in the second part follows a long term forecasting analysis in period of one year. Next section describes the data used for implementing and evaluating the forecasting models, as well as the pre-processing steps taken to transform the data into a suitable form. A detailed description of the implemented models and the various processing steps followed throughout the analysis, is also provided. Section IV presents the evaluation framework through which the various implemented models were tested, as well as the experimental results. Finally, in conclusion section, main contributions are reviewed and future directions are suggested.

MAIN PART

1. Time series modeling

Time series modeling and forecasting has fundamental importance to various practical domains. Time series modeling is a dynamic research area which has attracted attentions of researchers' community over the last few decades. The main aim of time series modeling is to carefully collect and rigorously study the past observations of a time series to develop an appropriate model which describes the inherent structure of the series. This model is then used to generate future values for the series, i.e. to make forecasts. Forecasting is a vital ingredient in the making of both long-term and short-term plans. When electricity sectors were regulated, utility monopolies used short-term load forecasts to ensure the reliability of supply and long-term demand forecasts as the basis for planning and investing in new capacity. *Short-term forecasting* generally involves horizons up to 1 hour ahead, as *Long-term forecasting*, with lead times measured in months, quarters or even years, concentrates on investment profitability analysis and planning, such as determining the future sites or fuel sources of power plants. In this study we implement both short term and long term forecasts.

1.1. Short-term forecasting

A forecasting analysis of period of 1 week is presented and the preprocessing of the data is demonstrated.

1.1.1. Data description

The data used in this study to implement the various forecasting models and evaluate their forecasting accuracy, were collected from a photovoltaic plant located in Crete, Greece. The dataset contains energy values from several panels of a photovoltaic plant during a total period of 7 days. The analyzed time period is from July 1, 2013, to July 7, 2013. The data granularity is 15 minutes, meaning that for each quarter of each day and photovoltaic plant, an energy value was available.

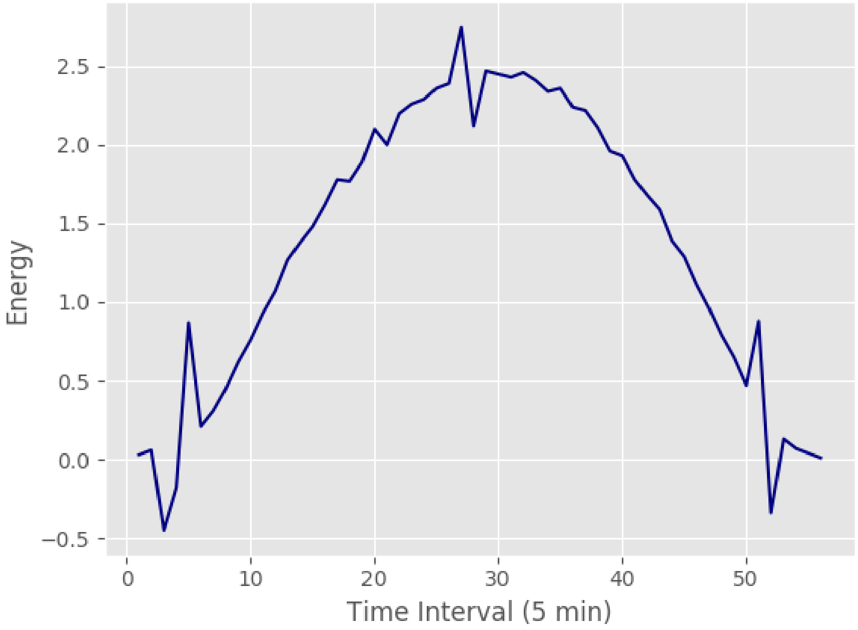


Figure 1: Daily power output from photovoltaic panel

Apart from the energy values, the dataset also contains values of the variables that describe the environmental state during the data gathering process. In particular, the dataset contains values for the solar irradiance, the ambient temperature and the panel's temperature. The granularity of these variables is the same as the one of the energy. This data is very useful, because, as it will be presented later, the meteorological variables are strongly correlated with the energy.

In table 1, we present the several variables that exist in the dataset and their measures.

Table 1: Data points with their description

Variable	Description	Measure
Timestamp	Local Hour	H
IntSolIrr	Solar Irradiance	W/m^2
TmpAmb	Ambient Temperature	^{o}C
TmpMdul	Temperature of photovoltaic panel	^{o}C
E-Total	Total Energy Output	kWh

1.1.2. Data Preprocessing

For the analysis, it was necessary to process the data by filling the missing data points. The values that are used are from dawn to dusk for each day. Each variable contains measurements for every 15 minutes. As it was already mentioned, the dataset contains values of several variables for each quarter of an overall period of seven days. Several quarters of the day were missing and we had to fill them using the method of interpolation. Firstly, the column containing the hour in-formation was completed. In order to fill the rest missing data points, *cubic spline interpolation* was used and it was preferred to linear interpolation due to better results.

Subsequently, another important step before analyzing and extracting results from the time series, is the outlier detection. Identification of outliers is very important in many fields that deal with time series analysis since they can contain information that may lead to an intervention of a process and prevent failures or abnormal operating conditions. An outlier can be defined as a data point in a time series that is significantly different from the rest of the data points. In this study two-sided median method for cleaning data was being used [[2]]. In this process, a neighborhood of points was defined and then the median of this neighborhood was calculated. If the absolute value of the difference between the point and the median is greater than a threshold then the point is an outlier and is replaced with the median. If it is smaller, then nothing happens.

In regard to make predictions, we had to produce time series from the original datasets. After filling the missing data points and removing the outliers, the next step was to construct time series from the original data. As already mentioned, the dataset contained energy data from 6 panels with similar characteristics located in the same photovoltaic park and meteorological data. We constructed one time series of energy values for each panel and day. Each time series has 56 values as the data granularity was 15 minutes and we used only the data points from dawn to dusk. In this way we constructed 42 time series in total for all the panels and examined days. Additionally, we constructed one time series of size 56 for each meteorological variable and examined day. Therefore, we had 21 time series in total for all meteorological variables and examined days.

1.2. Long-term forecasting

Besides the short-term forecasting analysis, a long-term forecasting completes this thesis. As in the short-term there is a data description and the preprocessing of the data is presented.

1.2.1. Data description

The data that we used in the second part of this study is a dataset covering a period of approximately one year. Specifically, from July, 1 2013 to July, 16 2014. The data granularity is also 15 minutes. As we mention before the dataset contains values for the solar irradiance, the ambient temperature and the panel’s temperature. These values are used also in this analysis.

1.2.2. Data Preprocessing

In order to proceed in the analysis, we had to fill the missing dataset and as we already mentioned we used cubic spline interpolation for this task. It was also necessary to detect the outliers and remove them. For this reason we implemented the two sided median method, as previously. For each day we had energy values from 6 PV panels and we calculated the mean energy value of each PV panel and then the average of these 6 means. As a result, we had the daily average value of energy of each day and with these values we constructed a time series as it is shown in figure 2.

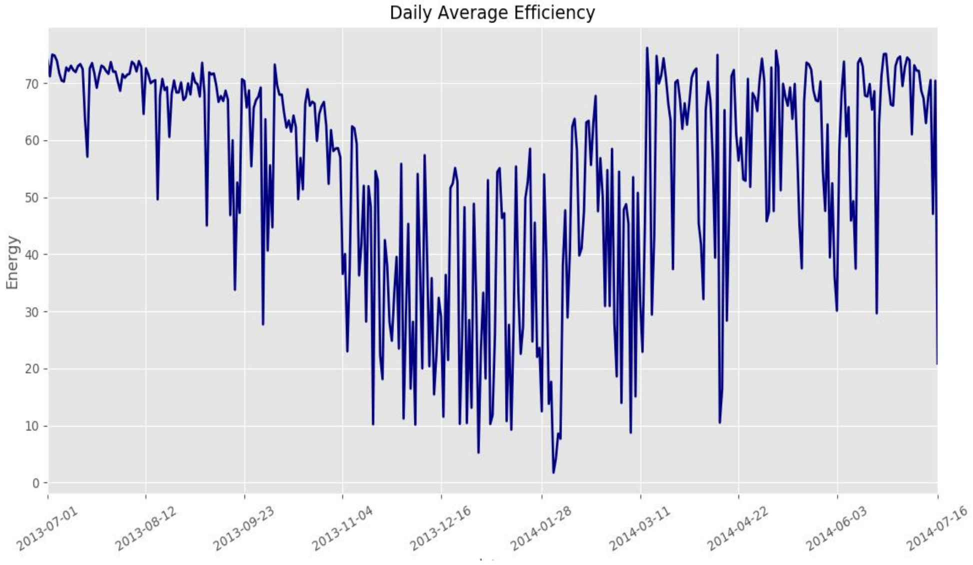


Figure 2: Daily Average Efficiency

In the previous figure we observe that our time series is noisy and a filter apart from helping us removing the noise, it would help us define the periods of “Summer” and “Winter” for the consequent analysis. As it is shown in Figure 3 the period of “Winter” is from November, 4 2013 to March, 13 2014. The rest of the time series constitutes the “Summer” period. We can observe the ‘Summer’ and the ‘Winter’ period in figures 4 and 5. The first day of the “Summer” period is July, 01 2013 and the last is July, 16 2014. There are several ways for cleaning a signal. We implemented a FFT filter which is based on the Fast Fourier Transform. An FFT Filter is a process that involves mapping a time signal from time-space to frequency-space in which frequency becomes an axis. By mapping to this space, we can get a better picture for how much of which frequency is in the original time signal and we can ultimately cut some of these frequencies out to remap back into time-space. Such filter types include low-pass, where lower frequencies are allowed to pass and higher ones get cut off, high-pass, where higher frequencies pass, and band-pass, which selects only a narrow range or "band" of frequencies to pass through. We used a low pass filter and specifically an order 3 low pass Butterworth filter.

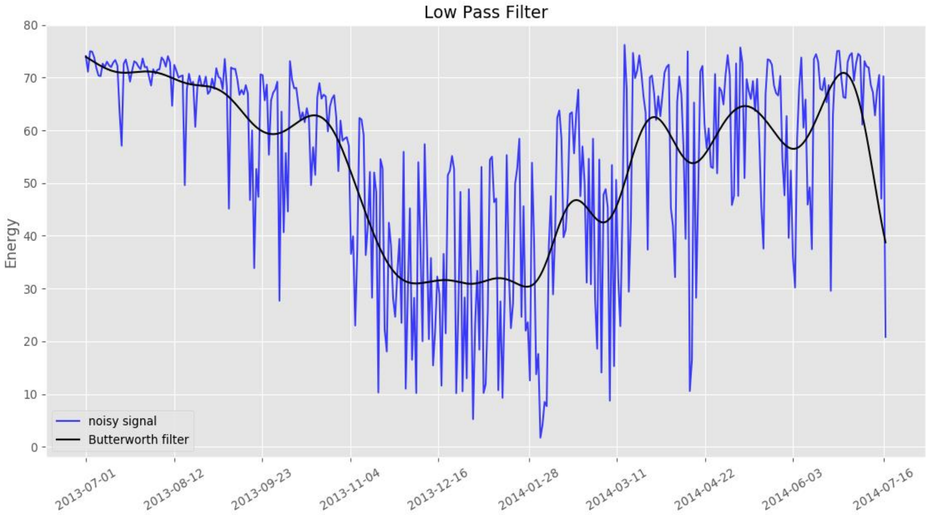


Figure 3: Daily Average Efficient with Butterworth Low Pass Filter of order 3

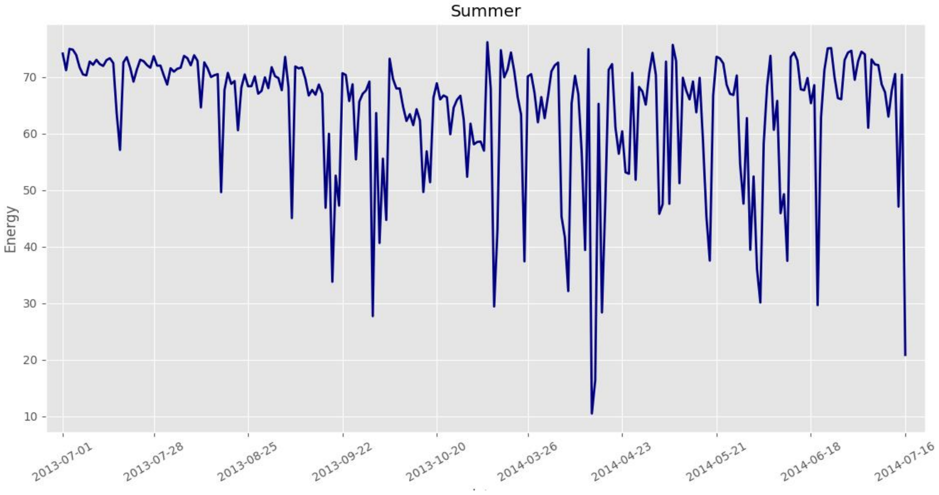


Figure 4: "Summer" period of daily average efficiency

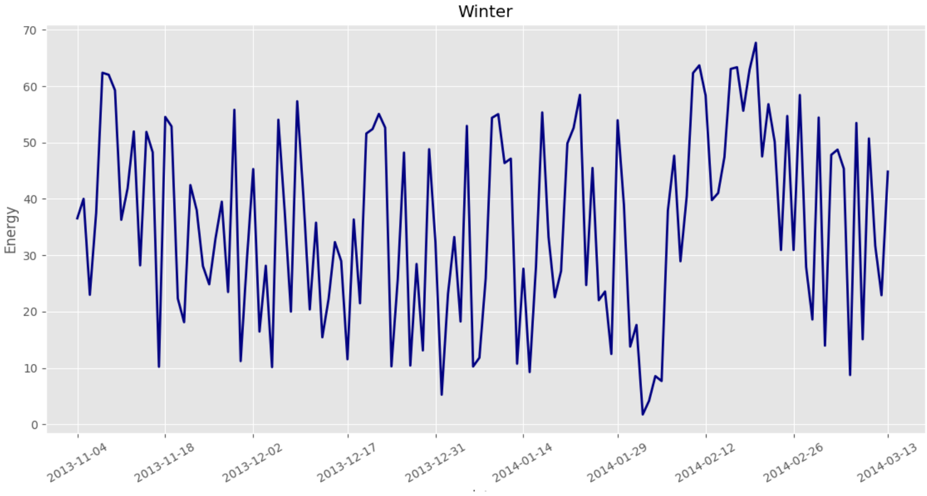


Figure 5: "Winter" period of daily average efficiency

2. Methodology

For the purpose of forecasting the power output several methods were implemented in this work for 1, 2, 3 and 4 quarters of time ahead for the analysis of short term forecasting and 1, 2, 3 and 4 days ahead for the analysis of the long term forecasting. The methods employed are:

- Autoregressive Integrated Moving Average (ARIMA).
- Artificial Neural Networks (ANNs).
- Improved Models which combine meteorological and power output data.

Time series forecasting can be conducted by several approaches. The more traditional methods are the linear ones such as moving average models, autoregressive models and simple linear regression models. Due to their lack of complexity in understanding and implementation, obtain the focus of many proposed works.

2.1. Autoregressive Integrated Moving Average (ARIMA)

A time series is a sequence of measurements of the same variable(s) made over time. Usually the measurements are made at evenly spaced intervals. In this work, the size of this interval is 15 minutes. ARIMA family is the most general class of linear models for forecasting a time series. An ARIMA (*p*, *d*, *q*) model consists of three parts. The autoregressive part (AR) with order *p*, the moving average part (MA) with order *q* and the integrated part (I) which describes the number of differencing steps (*d*) that should be taken in order to stationarize the time series. The model is defined by the following formula.

$$X_t = c + \varepsilon_t + \sum_{i=1}^{p} \varphi_i X_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} \tag{1}$$

where ε_t are the error terms that follow normal distribution with 0 mean and *c* arbitrary constant.

The auto-regressive part (AR) of the model has its origin in the theory that individual values of time series can be described by linear models based on preceding observations. The autoregressive processes have, in general, infinite non-zero autocorrelation coefficients that decay with the lag. The AR processes have a relatively "long" memory, since the current value of a series is correlated with all previous ones, although with decreasing coefficients. This property means that we can write an AR process as a linear function of all its innovations, with weights that tend to zero with the lag. The variables, which represent the new information that is added to the process at each instant, are known as innovations. The AR processes cannot represent short memory series, where the current value of the series is only correlated with a small number of previous values.

A family of processes that have this "very short memory" property are the moving average, or MA processes. The MA processes are a function of a finite, and generally small, number of its past innovations. The consideration leading to moving average models (MA models) is that time series values can be expressed as being dependent on the preceding estimation errors. Past estimation or forecasting errors are taken into account when estimating the next time series value.

2.1.1. Stationarity

A stationary time series is one whose statistical properties such as mean, variance, autocorrelation, etc. are all constant over time. The time series forecasting models can be implemented only on stationary time series. A stationary time series gives meaningful sample statistics such as means, variances, and correlations with other variables. In order to receive consistent, reliable results, the non-stationary time series needs to be transformed into stationary, before implementing the time series forecasting models. Non-stationary data, as a rule, are unpredictable and cannot be modeled or forecasted. In time series analysis several methods exist in order to check stationarity such as plotting the time series for visual inspection, calculation of summary statistics, like the mean or the variance of the observations in order to check for obvious or significant differences and the implementation of statistical tests to check if the expectations of stationarity are met or have been violated. For example in Figure 6 we can see the plot of a time series that it is stationary. Rolling mean does not change dramatically over time and as a result the time series is considered stationary.

There are several statistical tests to detect stationarity such as the Augmented Dicker Fuller test, the Phillips-Perron test which builds on the Dicker-Fuller test of null hypothesis, the Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test, the ADF-GLS test etc. Augmented Dicker-Fuller test is a type of statistical test called a unit root test. The intuition behind a unit root test is that it determines how strongly a time series is defined by a trend. Augmented Dicker-Fuller test uses an autoregressive model and optimizes an information criterion across multiple different lag values. The null hypothesis of the test is that the time series can be represented by a unit root, which is not stationary, has some time-dependent structure. The alternate hypothesis (rejecting the null hypothesis) is that the time series is stationary. Null Hypothesis (H0): If accepted, it suggests the time series has a unit root, meaning is non-stationary. Alternate Hypothesis (H1): The null hypothesis is rejected; it suggests the time series does not have a unit root, meaning it is stationary. The unit root test confirmed the result that we already knew, that the time series is not stationary. The null hypothesis was not rejected. A non-stationary time series can be transformed into a stationary by removing trend. In general, a systematic change in a time series that does not appear to be periodic is known as a trend. Trends can be applied to the whole time series and to parts or subsequences of a time series too.

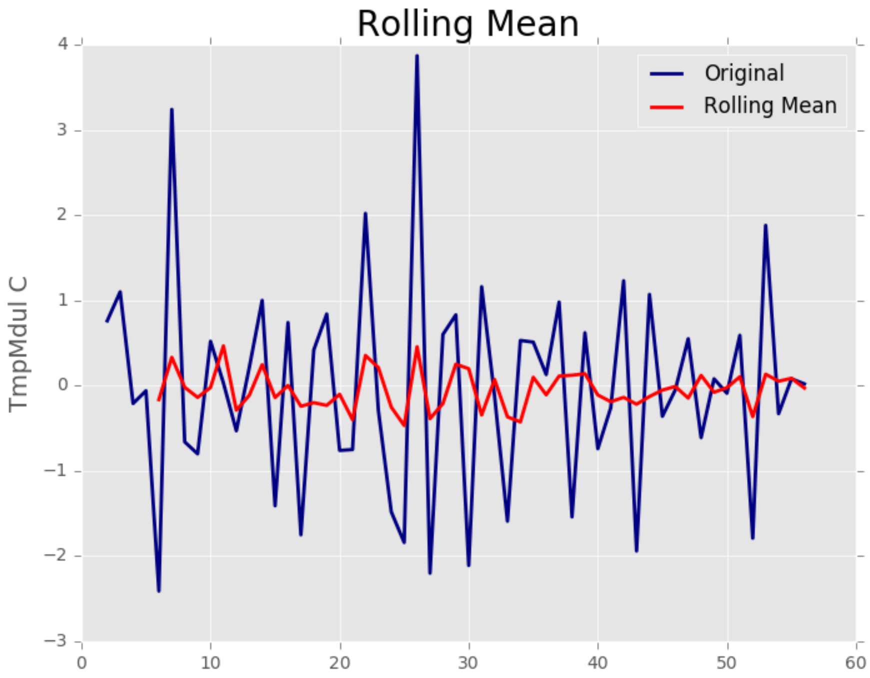


Figure 6: Rolling mean plot of Photovoltaic Panel's Temperature

Another statistical unit root test that is used, is the Phillips-Perron test which builds on the Dicker-Fuller test of null hypothesis. Davidson and MacKinnon (2004) report that the Phillips–Perron test performs worse in finite samples than the augmented Dickey–Fuller test. The Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test is another unit root test which is used in econometrics. Contrary to the most unit root tests, the presence of a unit root is not the null hypothesis but the alternative. Additionally, in the KPSS test, the absence of a unit root is not a proof of stationarity but, by design, of trend-stationarity. This is an important distinction since it is possible for a time series to be non-stationary, have no unit root yet be trend-stationary. KPSS-type tests are intended to complement unit root tests, such as the Dickey–Fuller tests. ADF-GLS test is a test for a unit root in an economic time series sample. It is used in statistics and econometrics and it was developed by Elliott, Rothenberg and Stock (ERS) in 1992 as a modification of the augmented Dickey–Fuller test (ADF).

2.1.2. Removing non-stationarity

For the purpose of making non-stationary time series stationary we conducted all the aforementioned methods. Through the process of plotting the original data, we detected obvious trend in the inputs values of energy. Trends can be applied to the whole time series and to parts or subsequences of a time series too. Meteorological variables showed that the mean and the variance remain constant over time from one period to the next. Moreover we split the data into two partitions and then we calculated the mean and the variance of each partition. The comparison of the summary statistics verified that the time series are non-stationary. As we mentioned before a method to check stationarity is the implementation of a statistical test. We used the Augmented Dicker-Fuller test which may be one of the more widely used. Concerning, short term analysis, as seen in Table 2, the statistical test indicates that the time series is non-stationary, for the critical values are smaller than the ADF value and hence the null hypothesis was accepted. One of the procedures in order to transform a time series into stationary time series, is differencing. Computing the differences between consecutive observations of a time series, is known as differencing. Occasionally the differenced data will not appear stationary and it may be necessary to difference the data a second time to obtain a stationary series. This is known as second order of differencing. In practice, it is almost never necessary to go beyond second-order differences. We used first and second order of differencing to achieve better results and differencing of order one was adequate for transforming our time series. Consequently, we ran the unit root again and we compared the critical values with the statistical value of the test and the null hypothesis was rejected and as a result the time series is stationary. Another transformation is the logarithm transformation. In this method, firstly the logarithm of each observation is calculated and then, the differences of the consecutive log-observations of a time series are computed. The log transformation for this research did not give us better results and for that reason it was rejected as a method of making times series stationary.

Table 2: Results of Augmented-Dicker Fuller statistical test

	Irradiance	Temperature	Energy
ADF statistics value	-0.878	-3.406	-2.133
Critical Value: 5%	-2.918	-2.917	-2.923
Critical Value: 10%	-2.597	-2.596	-2.599
Critical Value: 1%	-3.560	-3.558	-3.571

The results of ADF test after differencing are shown in Table 3.

Table 3: Results of Augmented-Dicker Fuller statistical test after stationarize the time series

	Irradiance	Temperature	Energy
ADF statistics value	-7.934	-7.336	-4.203
Critical Value: 5%	-2.918	-2.917	-2.923
Critical Value: 10%	-2.597	-2.596	-2.599
Critical Value: 1%	-3.560	-3.558	-3.571

On the other hand, in the case of long term analysis, we observe from Table 4 and Table 5 that there is no need of differencing, detrending or log transformation to proceed with the analysis.

Table 4: Results of Augmented Dicker-Fuller statistical test of the 'Summer' period

Summer	Irradiance	Temperature	Energy
ADF statistics value	-8.123	-4.119	-7.626
Critical Value: 5%	-2.873	-2.874	-2.873
Critical Value: 10%	-2.573	-2.571	-2.573
Critical Value: 1%	-3.457	-3.458	-3.457

Table 5: Results of Augmented Dicker-Fuller statistical test of the 'Winter' period

Winter	**Irradiance**	**Temperature**	**Energy**
ADF statistics value	-4.276	-3.904	-5.975
Critical Value: 5%	-2.885	-2.884	-2.884
Critical Value: 10%	-2.579	-2.579	-2.579
Critical Value: 1%	-3.483	-3.482	-3.483

2.1.3. ACF and PACF

In order to identify which linear model is suitable for forecasting in our time series, we use the Autocorrelation Function (ACF) and the Partial Autocorrelation Function (PACF) plots. Plotting the Autocorrelation function (ACF) plot and the Partial Autocorrelation function plot (PACF) gives an idea of which lag variables may be good candidates for use in a predictive model. Moreover, gives an idea of how the relationship between the observation and its historic values changes over time and can be used for the following two purposes: Firstly, to detect non-randomness in data and secondly, to identify an appropriate time series model if the data are not random. An autocorrelation plot shows the value of the autocorrelation function (acf) on the vertical axis. It can range from –1 to 1. The horizontal axis of an autocorrelation plot shows the size of the lag between the elements of the time series. The autocorrelation with lag zero always equals 1, because this represents the autocorrelation between each term and itself. A partial autocorrelation is a summary of the relationship between an observation in a time series with observations at prior time steps with the relationships of intervening observations removed.

Autocorrelation function is one of the tools used to find patterns in the data. Specifically, the autocorrelation function tells us the correlation between points separated by various time lags. In general, if ACF has an exponentially decreasing appearance and PACF becomes zero at specific lag p, then an Autoregressive model (AR (p)) is suitable for forecasting. On the other hand, if ACF becomes zero in lag p and PACF decreases exponentially, a Moving Average (MA (p)) should be used.

The ACF and PACF plots for energy time series that correspond for a specific photovoltaic panel at July 1, 2013 are shown in Figure 7. From their form we understand that an AR model with order p = 1 (AR(1)) is suitable for representing the data and therefore for forecasting. The same result applies to all time series of our dataset.

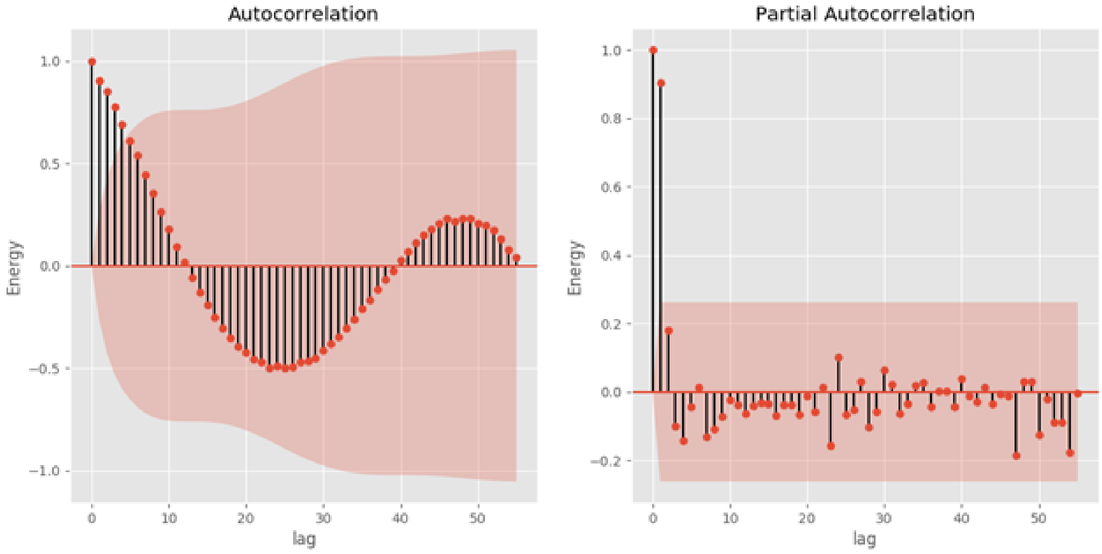


Figure 7: ACF and PACF plots for model identification

In Figure 8, the ACF and PACF plots for energy time series from July, 1, 2013 to July, 16, 2014 are being demonstrated. It is shown that an AR model is appropriate for our analysis with orders 1, 2 or 3. After experimentation we concluded that the AR model of order 2 is the one that leads to the best forecast accuracy.

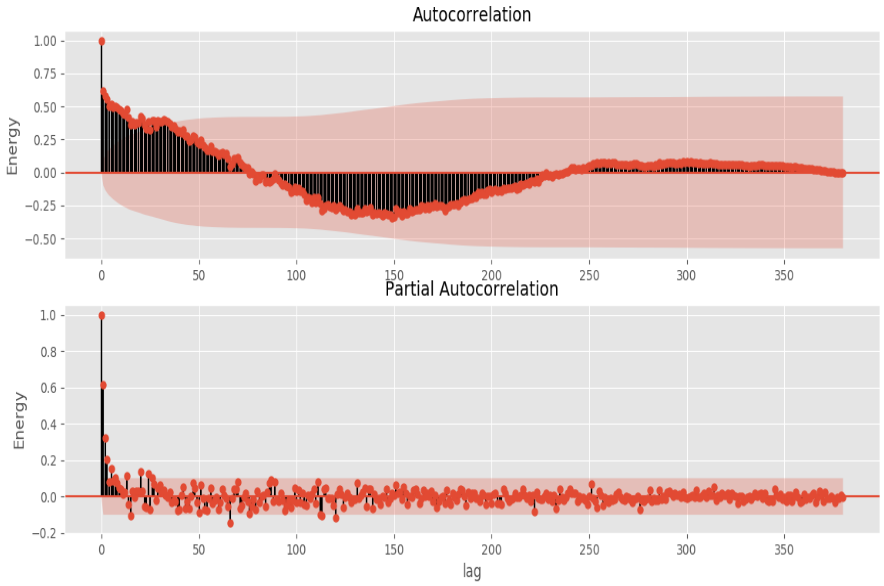


Figure 8: ACF and PACF plots for model identification of daily average efficiency

2.1.4. Implementation of ARIMA models

As already mentioned, for short term forecasting, the number of differencing steps taken to stationarize our time series is 1, i.e. *d = 1*. Also, as indicated by the ACF and PACF plots, the ARIMA model that fits more in our time series has strong autoregressive part with order *p = 1*, and no moving average part, i.e. *q = 0*. Consequently, the specific linear model used for power output forecasting in PV systems was ARIMA (1,1,0). We also considered the ARIMA (2,1,0) and ARIMA (3,1,0) models for a more meticulous research and also for detecting if the choice of the order of the autoregressive part through the ACF and PACF was correct. For ARIMA (1,1,0) model (1) becomes:

$$X_t = c + X_{t-1} + \varphi_1 (X_{t-1} - X_{t-2}) \qquad (2)$$

The training process of an ARIMA model consists of the estimation of the φ and θ parameters. Using the available data, a set of equations like (2) are formulated with unknowns the φ and θ parameters. The number of equations in this set is far greater than the number of the parameters. Therefore, this *over-determined system* is not solved analytically but in the least-squares way. In this work, we had to estimate 1 parameter for the ARIMA (1,1,0) model and 2 and 3 parameters for the ARIMA (2,1,0) and ARIMA (3,1,0) models. We used the *Singular Value Decomposition (SVD)* method to solve the formulated overdetermined systems and estimate the corresponding parameters.

Regarding long term forecasting analysis, the linear model for power output forecasting in PV systems is an ARIMA (1,0,0). As we considered in short term forecasting analysis the ARIMA with order 2 and 3 for a more scrupulous research, thus we considered ARIMA (2,0,0) and ARIMA (3,0,0). For ARIMA (1,0,0) the model equation is:

$$X_t = c + \varphi_1 X_{t-1} \qquad (3)$$

2.2.Artificial Neural Network (ANN)

Apart from linear methods, extensively studies were conducted using non-linear methods in time series forecasting. Many real life time series that need to be forecasted cannot be implemented with linear methods as they are rarely pure linear. One of the most widely used non-linear models, due to its flexibility in non-linear model capability [[20]], is artificial neural networks.

Artificial Neuron is a computational model inspired in the natural neurons. Natural neurons receive signals and when the signal is strong enough, the neuron is activated. An artificial neural network consists of several connected layers. Layers are made up of a number of interconnected '*nodes*' which contain an *'activation function'*. Patterns are presented to the network via the *'input layer'*, which communicates to one or more *'hidden layers'* where the actual processing is done via a system of weighted *'connections*'. The hidden layers then link to an *'output layer'* where the answer is output. Figure 9 depicts a general topology of an artificial neural network. The signals travels from input to output and are real numbers between 0 and 1. A Multilayer Perceptron (MLP) is a feedforward artificial neural network model that maps sets of input data onto a set of appropriate outputs. An MLP consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one.

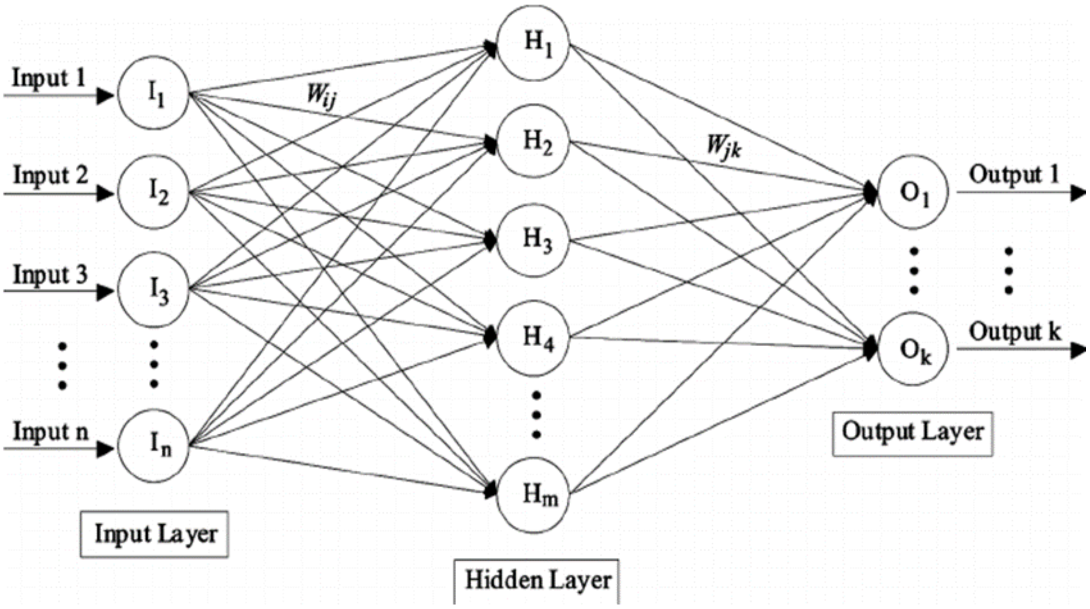


Figure 9: General topology of ANN

Except for the input nodes, each node is a neuron with a nonlinear activation function. The activation function of a node defines the output of that node given an input or a set of inputs. Moreover, it is a logistic regressor where instead of feeding the input to the logistic regression, inserts an intermediate layer, called the hidden layer, that has a nonlinear activation function, usually hyperbolic tangent or sigmoid.

A sigmoid function is a mathematical function having an "S" shaped curve (sigmoid curve). Often, sigmoid function refers to the special case of the logistic function and it is defined by the formula.

$$S = \frac{1}{1+e^{-t}} \tag{4}$$

In addition, sigmoid functions have finite limits at negative infinity and infinity, most often going either from 0 to 1 or from −1 to 1, depending on convention.

Another activation function, as previously mentioned is the hyperbolic tangent. Hyperbolic tangent is the solution to the differential equation

$$f' = 1 - f^2 \tag{5}$$

With

$$f(0) = 0 \tag{6}$$

and the nonlinear boundary value problem:

$$\frac{1}{2}f'' = f^3 - f;\ f'(\infty) = 0 \tag{7}$$

An identity function is a function that always returns the same value that was used as its argument and it can also be used as an activation function. In equations, the function is given by

$$f(x) = x \tag{8}$$

In the context of artificial neural networks, the rectifier is an activation function too an it is defined as:

$$f(x) = \max(0, x) \tag{9}$$

where x is the input to a neuron. The rectifier is, as of 2015, the most popular activation function for deep neural networks.

Usually, the ANNs utilize a supervised learning technique called *backpropagation* for training, i.e. estimating the values of the weights. Backpropagation is a common method of training ANNs and is used in conjunction with optimization methods such as gradient descent. The algorithm repeats a two phase cycle, propagation and weight update.

The principle of the backpropagation approach is to model a given function by modifying internal weightings of input signals to produce an expected output signal. The system is trained using a supervised learning method, where the error between the system's output and a known expected output is presented to the system and used to modify its internal state. The output of the network is compared to the desired output, using a loss function, and an error value is calculated for each of the neurons in the output layer. The error values are then propagated backwards, starting from the output, until each neuron has an associated error value which roughly represents its contribution to the original output.

Technically, the backpropagation algorithm is a method for training the weights in a multilayer feed-forward neural network. As such, it requires a network structure to be defined of one or more layers where one layer is fully connected to the next layer. A standard network structure is 1 input layer, 1 hidden layer, and 1 output layer.

The importance of this process is that, as the network is trained, the neurons in the intermediate layers organize themselves in such a way that the different neurons learn to recognize different characteristics of the total input space. After training, when an arbitrary input pattern is present which contains noise or is incomplete, neurons in the hidden layer of the network will respond with an active output if the new input contains a pattern that resembles a feature that the individual neurons have learned to recognize during their training.

2.2.1. Implementation of ANN

In the short-term forecasting analysis, we implemented feed-forward ANNs with, 1 input, 1 hidden and 1 output layer for forecasting power output in PV systems. The number of neurons in the input layer was selected equal to the order of the evaluated

ARIMA models (i.e. 1, 2 and 3), while in the output layer we had only 1 neuron. Furthermore, Table 6 shows that we can reject the null hypothesis of the Wilcoxon test as the p-values are not greater than 0.05. Hence, in the choice of the number of the neurons in the input layer there is a significant difference. In order to select the number of neurons in the hidden layer and the appropriate activation function, we ran simulations with different numbers of nodes (from 1 to 100) and different activation functions. The activation functions tested were the logistic sigmoid function, the hyperbolic tangent function and the rectified linear unit function. At each simulation cycle, we trained a different ANN and used it to make forecasts on a specific test time series. The ANN that yielded the best results in terms of forecasting accuracy was the one with 3 neurons in the hidden layer and the logistic sigmoid function as activation function for all the neurons, apart from the neuron in the output layer for which the identity function was used. Three different variations of this ANN (with 1, 2 and 3 neurons in the input layer) were implemented and compared with the corresponding ARIMA models.

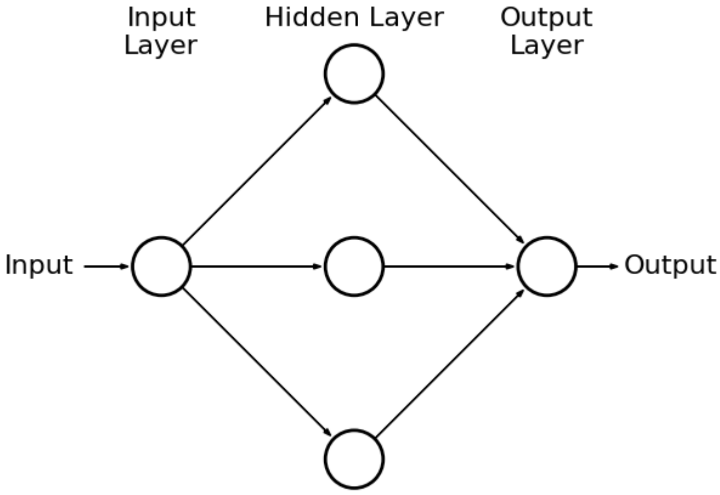


Figure 10: Topology of implemented ANN with 1 input, 3 hidden and 1 output nodes.

Table 6: Wilcoxon Statistical Test's Results

	ANN – 1 neuron	**ANN – 2 neurons**	**ANN – 3 neurons**
Wilcoxon p-value	1.02e-05	5.50e-06	5.36e-06

On the other hand, regarding the long-term analysis, we conducted the same statistical test in order to confirm the significant difference in the input layer's neurons as we can observe in the following tables (Table 7 and Table 8) and we ran the same simulations. For the 'Summer' period the best ANN topology is the one with 1 neuron in the input layer, 20 neurons in the hidden layer and the hyperbolic tangent function as activation function and 1 neuron in the output layer. As for the 'Winter' period is the one with 1 neuron in the input layer, 20 neurons in the hidden layer and the logistic sigmoid function as activation function and 1 neuron in the output layer. In the output layer the identity function was used in both periods.

Table 7: Results of Wilcoxon Statistical Test – 'Summer' period

Summer	**ANN – 1 neuron**	**ANN – 2 neurons**	**ANN – 3 neurons**
Wilcoxon p-value	7.56e-10	1.11e-09	1.63e-09

Table 8: Results of Wilcoxon Statistical Test – 'Winter' period

Winter	**ANN – 1 neuron**	**ANN – 2 neurons**	**ANN – 3 neurons**
Wilcoxon p-value	1.23e-05	1.82e-05	2.70e-05

One very common problem that arises when training ANNs, and especially ANNs with large number of layers and neurons, is *overfitting*. Overfitting is the case where the ANN performs very well on the training data, but poorly on the previous unseen test data, in other words fails to generalize well on new data. There are several ways to overcome overfitting, e.g. increasing the number of training instances or reducing the number of layers and neurons and thus the number of unknown parameters. These methods are not usually adopted reluctantly by the researchers because on one hand the acquisition of more training samples is a costly process, and on the other the reduction of the number of layers and neurons may lead to less powerful networks. Fortunately, there is another way for overcoming overfitting which is called regularization. In this, an extra term, which takes into account the magnitude of the weights of the network, is added to the cost function of the training algorithm. Usually, this term is the sum of the squares of all the weights of the network (excluding the biases) multiplied by a constant called *regularization parameter*. After training the network using the updated cost function, we can see that the effect of the overfitting is reduced. As will be seen at the evaluation framework, in this work we considered only the test error of our implemented ANN models and not the training error, and therefore we did not examine in depth the possibility of arising the problem of overfitting.

2.3.Improved models

Renewable energy sources such as solar energy, are weather dependent and it is evident that there is a sort of relationship regarding the forecast. In order to measure the dependency between the meteorological data and the energy production data we calculated *Pearson’s correlation coefficient* and *Spearman’s correlation coefficient*. A major goal of the examination of the improved models is to evaluate the performance of the models and to testify the existence of a crucial role in the accuracy by these models. Oudjana et al. [[5]] made power predictions based on weather variables correlations with the PV power output and they achieved better models. The importance of implementation of improved models for power output forecasting lies in the fact of minimizing the error and achieving even better results.

2.3.1. Correlations

Correlation is a bivariate analysis that measures the strengths of association between 2 variables and the direction of the relationship. In terms of the strength of relationship, the value of the correlation coefficient varies between -1 and +1. When the value of the correlation coefficient lies around ± 1, then it is said to be a perfect degree of association between the 2 variables. As the correlation coefficient value goes towards 0, the relationship between the 2 variables will be weaker. The direction of the relationship is simply the + (indicating a positive relationship between the variables) or - (indicating a negative relationship between the variables) sign of the correlation. In statistics, there are 4 ordinary, different ways to measure correlation: Pearson correlation, Kendall rank correlation, Spearman correlation, and the Point-Biserial correlation.

Pearson’s correlation coefficient is a statistic measuring the linear interdependence between 2 variables or two sets of data. The value of the coefficient ranges from -1 to +1, with +1 indicating a perfect positive linear relationship, -1 indicating a perfect negative relationship and 0 showing none existing relationship. Correlation factor +1 means a perfect correlation between the 2 variables. In other words, a scatter plot of the 2 variables will show that all points fit perfectly into a straight line. Coefficient 0 means that the points in the spreadsheet are randomly distributed around any straight line designed or are arranged to approach a curve. In general, sign "+" means positive correlation which means the values of one variable increase according to the values of the other and sign "-" means negative correlation where the values of a variable decrease according to the value of the other. The formula for calculating the coefficient for 2 variables *X* and *Y* is the following:

$$\rho = \frac{\text{cov}(X, Y)}{\sigma_x \sigma_y} \tag{10}$$

where *cov(X,Y)* the covariance and σ_x, σ_y the standard deviations of *X* and *Y* respectively.

Spearman’s correlation coefficient is a statistical measure of the strength of a monotonic relationship between paired data. The value of the coefficient ranges from -1 to +1, with +1 or -1 indicating that each of the variables is a perfect monotone function of the other. If Y tends to increase when X increases, the Spearman correlation coefficient is positive. If Y tends to decrease when X increases, the Spearman correlation coefficient is negative. A zero Spearman correlation shows that there is no tendency for Y to either increase or decrease when X increases. Spearman's association increases in size when X and Y are closer to being perfect monotonous functions of each other. When X and Y have an absolute monotonic relationship, the Spearman correlation coefficient becomes +1. The Spearman correlation coefficient is defined as the Pearson’s correlation coefficient between the corresponding *ranked* variables. The formula of Spearman’s coefficient is the following.

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 (y_i - \bar{y})^2}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{11}$$

where x_i, y_i are the ranks of X_i and Y_i respectively.

In cases that there are not equal values, a more simple procedure is used with the use of the following formula.

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} \tag{12}$$

where d_i is the difference between the ranges of each observation of the 2 variables.

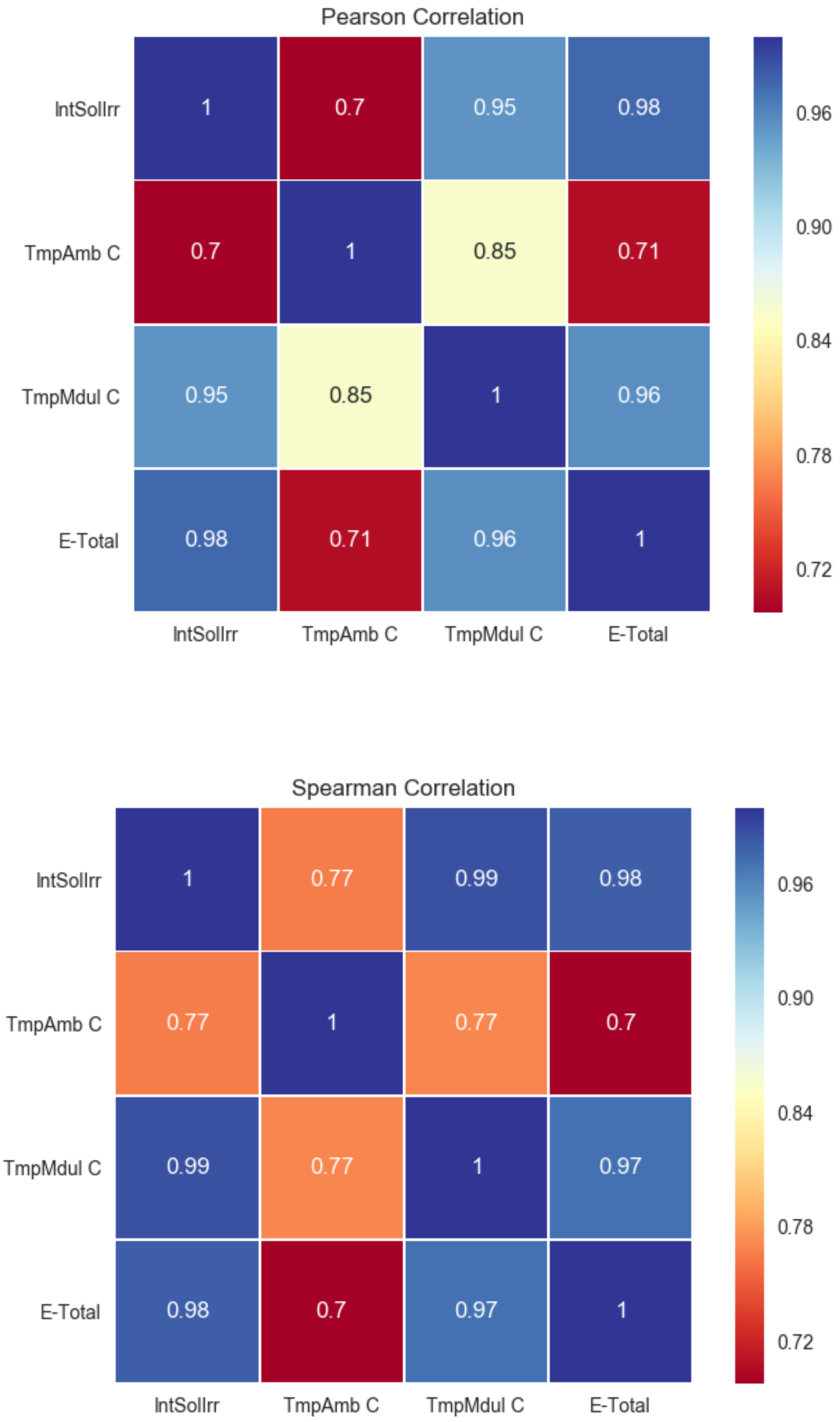


Figure 11: Heatmap of Pearson's correlation (above) and Spearman's correlation (below) between energy (E-Total), solar irradiance (IntSolIrr), ambient temperature (TmpAmb C) and panel's temperature (TmpMdul C).

The Pearson's and Spearman's coefficient values between the energy and the meteorological variables are shown in Figure 11. As shown in the last row of both matrices, the coefficients values between energy and the solar irradiance and panel's temperature are very close to 1 (0.98 and 0.96 respectively for Pearson's coefficient and 0.98 and 0.97 for Spearman's) and between energy and ambient temperature is approximately 0.7 (0.71 for Pearson's coefficient and 0.7 for Spearman's). These results indicate strong linear correlation between energy and the meteorological variable. Based on this conclusion, we enhanced the implemented forecasting models (both ARIMA and ANN) by adding the meteorological variables as explanatory variables of the models. Specifically, we implemented one ARIMA model with explanatory variables the energy and the solar irradiance, one with the energy and the panel's temperature and one with the energy and the ambient temperature. Similarly, for the case of the ANN model. Subsequently, we have 6 *enhanced* models in addition to the initial *base* models.

2.3.2. Improved ARIMA models

As in the case of base models, we implemented enhanced ARIMA models of order 1, 2 and 3 too. The equation that describes the new, enhanced ARIMA(1,1,0) model is:

$$X_t = c + X_{t-1} + \varphi_1(X_{t-1} - X_{t-2}) + \varphi_2(Y_{t-1} - Y_{t-2}) \tag{13}$$

And the improved ARIMA (1,0,0) model, of the long term analysis, is described by the formula:

$$X_t = c + \varphi_1 X_{t-1} + \varphi_2 Y_{t-1} \tag{14}$$

The equations are not solved analytically but in the least-squares way.

2.3.3. Improved ANN models

The improved ANN models were implemented based on the ANN of Section 2 but they have a different topology with 2 inputs, 3 hidden and 1 output nodes, as shown in Figure 12. Following the above procedure, different nodes in the hidden layer. In particular, we ran simulations from 1 to 100 nodes and with also distinct activation functions. The topology of ANN which performs better is the topology shown in Figure 12.

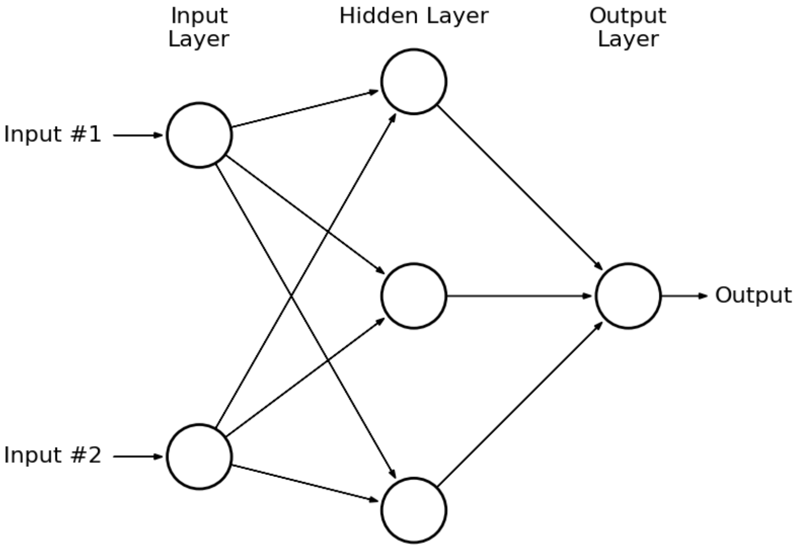


Figure 12: ANN topology of improved model

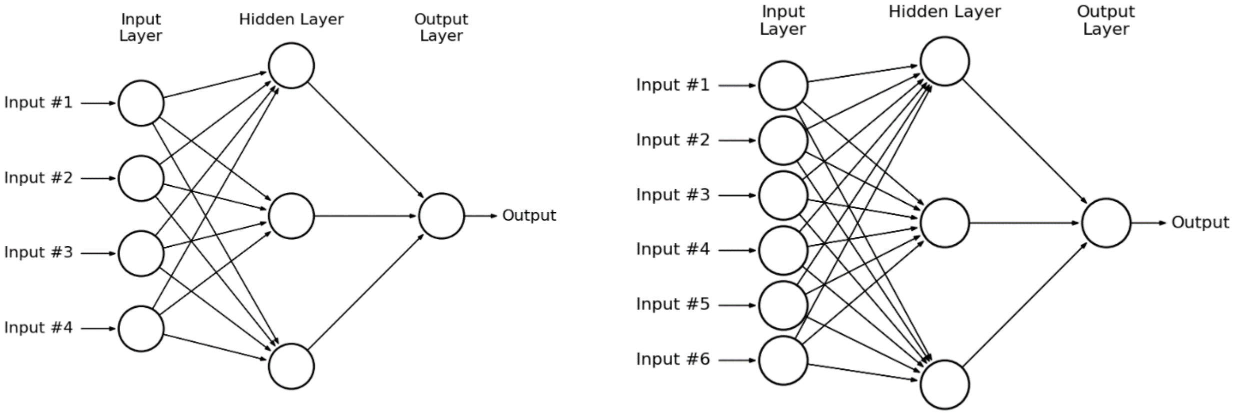


Figure 13: ANN topology order 2 (left) and order 3 (right) of the improved models

In the case of the long term analysis, we also ran simulations with different (from 1 to 100) nodes in the hidden layer and also distinct activation functions. The topology of the improved ANN models differs among the enhanced models. The number of input and output neurons are the same for all the improved models but the number of the neurons in the hidden layer tend to differ. Specifically, in the improved model with explanatory variables the energy and the solar irradiance in the "Summer" period the ANN topology is, 2 neurons in the input layer, 20 neurons in the hidden with rectified linear unit activation function and 1 neuron in the output layer. In the improved model with explanatory variables the energy and the ambient temperature in the "Summer" period the ANN topology is, 2 neurons in the input layer, 30 neurons in the hidden with rectified linear unit activation function and 1 neuron in the output layer. As for the improved model with explanatory variables the energy and the panel's temperature in the "Summer" period the ANN topology is, 2 neurons in the input layer, 90 neurons in the hidden with logistic sigmoid activation function and 1 neuron in the output layer. Regarding the "Winter" period, in the improved model with explanatory variables the energy and the solar irradiance the ANN topology is, 2 neurons in the input layer, 20 neurons in the hidden with logistic sigmoid activation function and 1 neuron in the output layer. In the improved model with explanatory variables the energy and the ambient temperature the ANN topology is, 2 neurons in the input layer, 50 neurons in the hidden with logistic sigmoid activation function and 1 neuron in the output layer. Finally, in the improved model with explanatory variables the energy and the panel's temperature the ANN topology is, 2 neurons in the input layer, 50 neurons in the hidden with hyperbolic tangent activation function and 1 neuron in the output layer. For all the improved models the identity activation function was used in the output layer. The number of the neurons in the input layer of the ANN model of order 2, is 4 neurons and in the ANN model of order 3, 6 neurons.

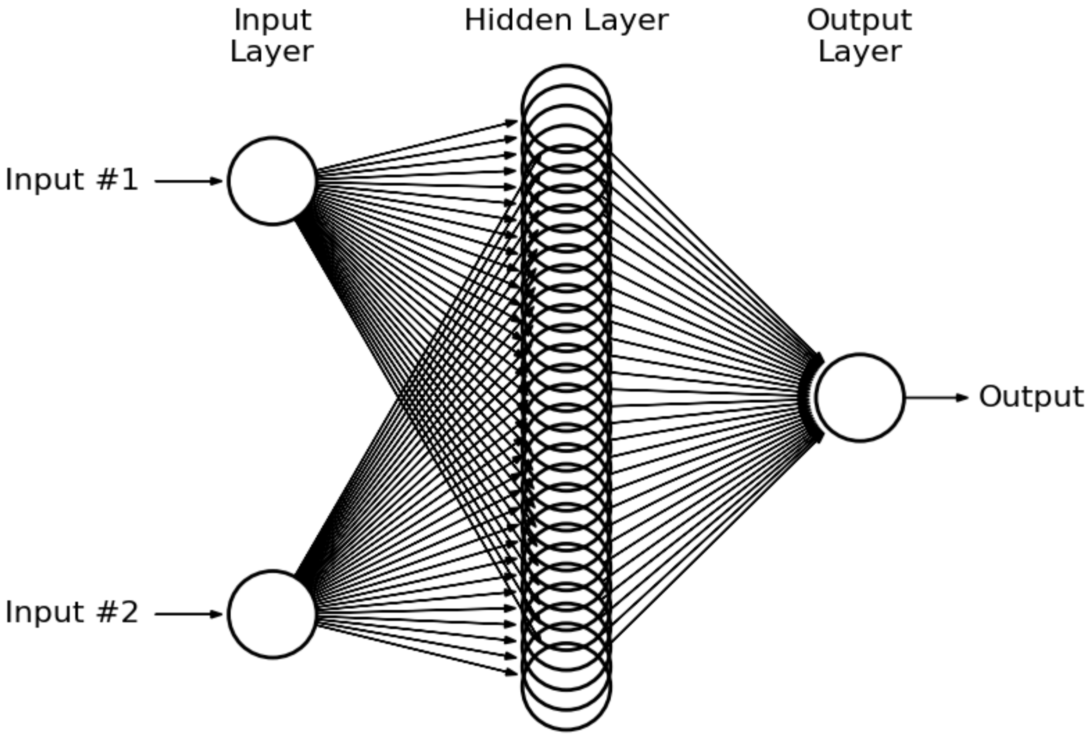


Figure 14: ANN Topology of improved model of order 1 in long term analysis (energy + solar irradiance as explanatory variables)

3. Experiments Tools

The presented models were developed in the high-level programming language Python (version 3). Python was first released in 1991 by Guido van Rossum, who named the language after the BBC show "Monty Python's Flying Circus". It can be used for everything from web development to software development and scientific applications. Python is one of those rare languages which can claim to be both *simple* and *powerful*. Python's elegant syntax and dynamic typing, together with its interpreted nature, make it an ideal language for scripting and rapid application development in many areas on most platforms. With the term interpreted we mean that Python does not need compilation to binary. You just *run* the program directly from the source code. This pseudo-code nature of Python is one of its greatest strengths. It allows you to concentrate on the solution to the problem rather than the language itself.

Additionally, the Anaconda set of packages for data science was used. Anaconda is the leading open data science platform powered by Python. The open source version of Anaconda is a high performance distribution of Python and R and includes over 100 of the most popular Python, R and Scala packages for data science.

The desktop computer that was used has Windows 10 Professional, 64bit operating system, a 6 core processor at 3.5GHz and also 8GB RAM memory.

4. Results

In order to implement and evaluate our models, we split the set of our time series into training and test subsets (80% in the training and 20% in test). We used the training time series to build our models and the test to make forecasts. The forecasts were made for 1 to 4 steps (15-minute intervals in case of short term forecasting and 1 day interval in case of long term forecasting) ahead in time. We repeated this process in all possible ways for the purpose of cross-validating the results. We used all training and test datasets for each implemented model (both base and enhanced) and we compared the results in terms of forecasting accuracy using the *Symmetric Mean Absolute Percentage Error (SMAPE)* metric, which is defined by the following formula:

$$SMAPE = \frac{100\%}{n}\sum_{t=1}^{n}\frac{|F_t - A_t|}{|A_t| + |F_t|} \tag{15}$$

where F_t is the forecasted value and A_t the actual value. We selected SMAPE over other error metrics because it provides a result between 0% and 100% which is more easily interpretable. The results of the experiments are presented in the following section.

4.1. Short-Term Forecasting

The original dataset had cumulative power outputs and we made forecasts firstly in the cumulative data points. In order to make this time series stationary we had to use the method of detrending instead of the method of differencing as in the rest of the dataset. As we can observe in Figure 15 ARIMA(1,1,0) fits the data better than the ANN model with 1 input. The predicted values are almost equal to the real ones and SMAPE results, (Table 9), confirms it.

Table 9: SMAPE results of ARIMA and ANN of cumulative time series

	ARIMA (1,1,0)	ANN - 1 neuron
SMAPE (%)	0.0006058	0.0010503

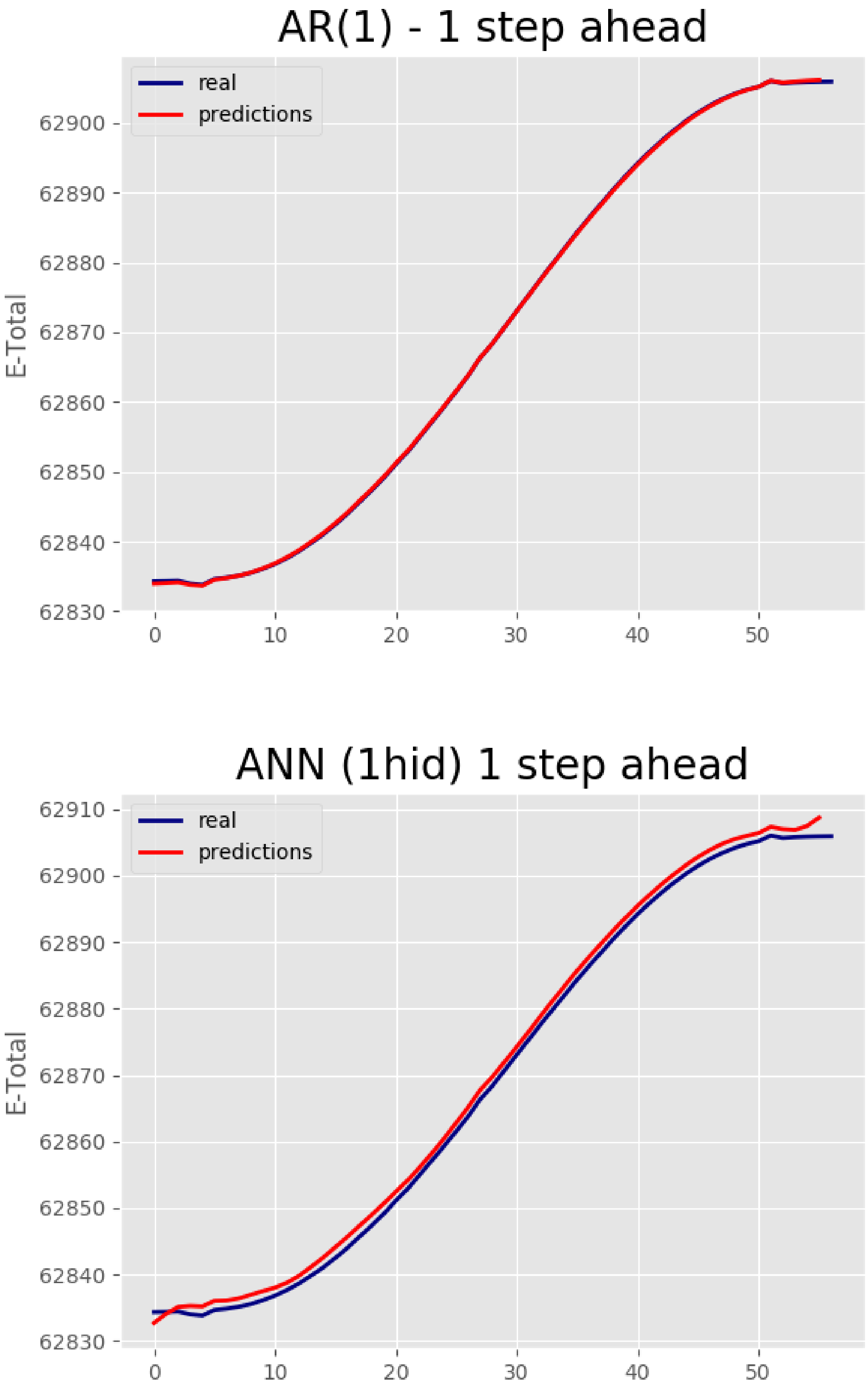
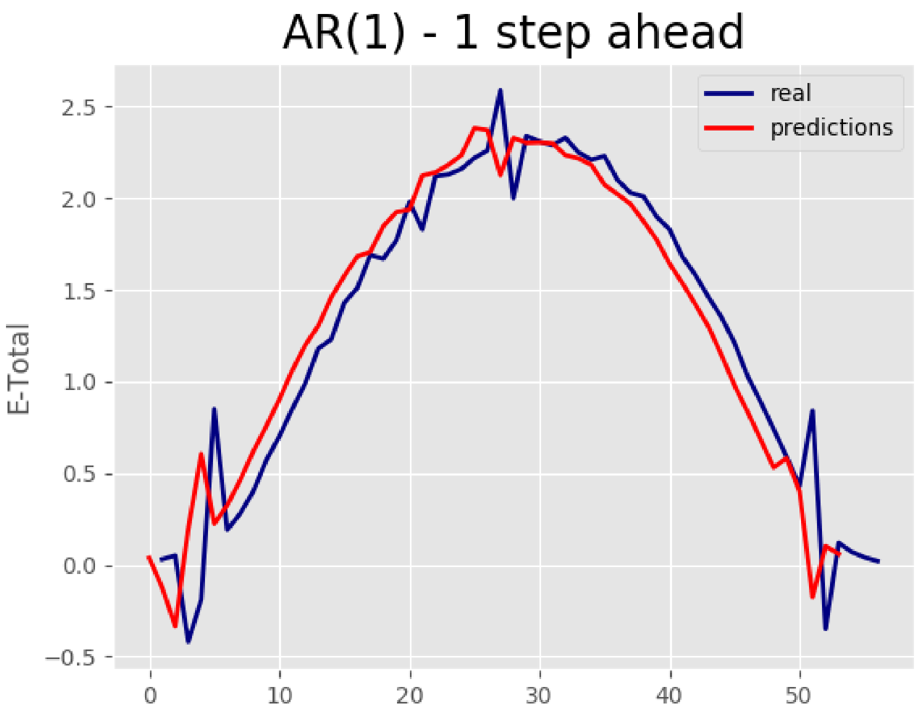
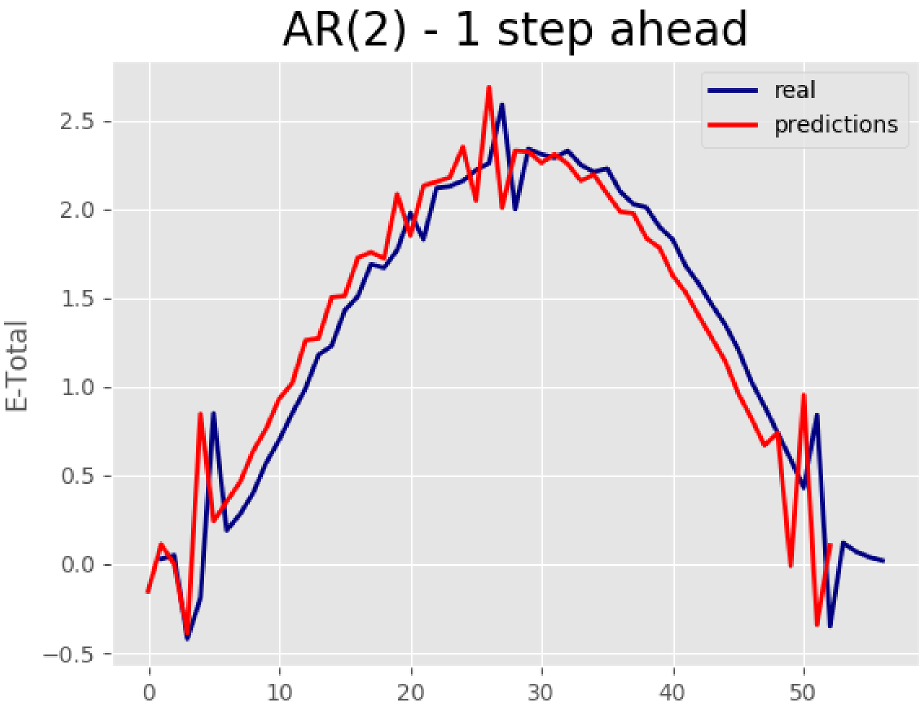


Figure 15: ARIMA and ANN fitting in cumulative time series of energy

As it is shown in Figure 16, the model that performs better among the linear models is ARIMA(1,1,0). This result is also demonstrated in Table 10, where we can see the average forecasting results for all steps ahead of the implemented ARIMA models that use only the energy as explanatory variable.





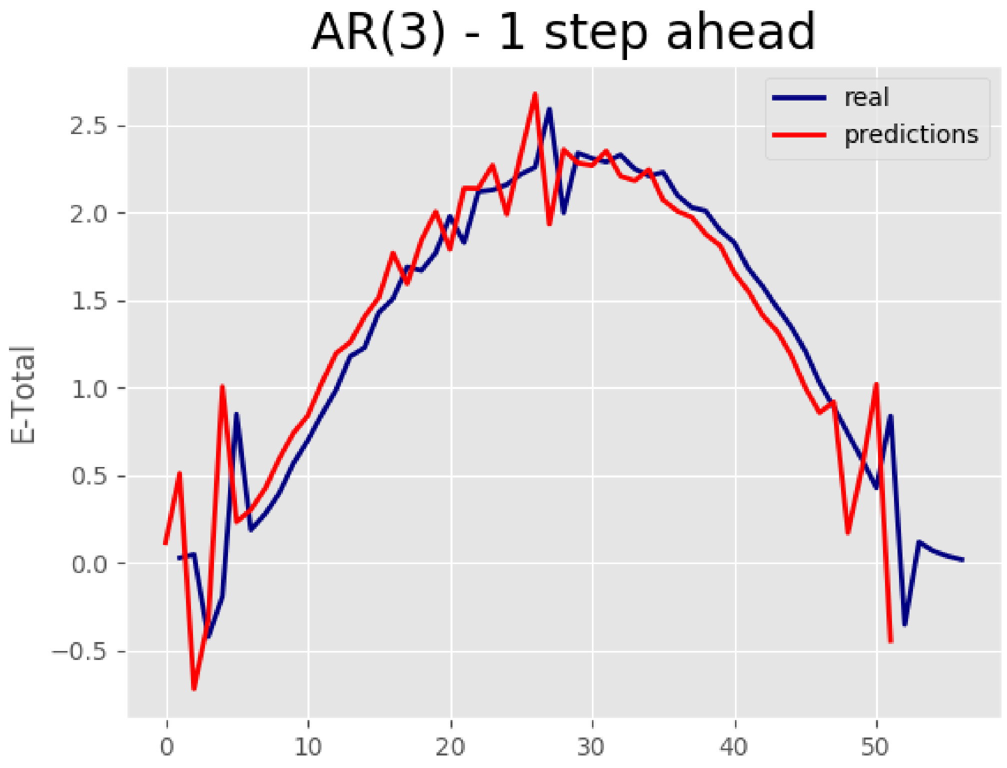


Figure 16: Plots of ARIMA models fitting for one step ahead

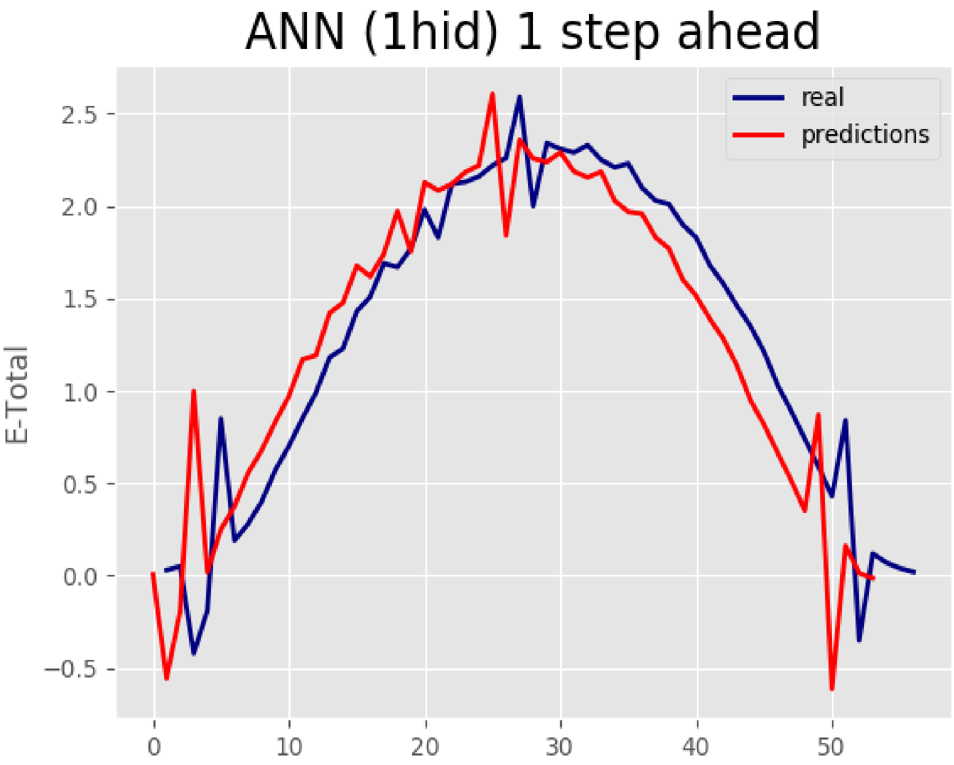
Table 10: SMAPE results of ARIMA models

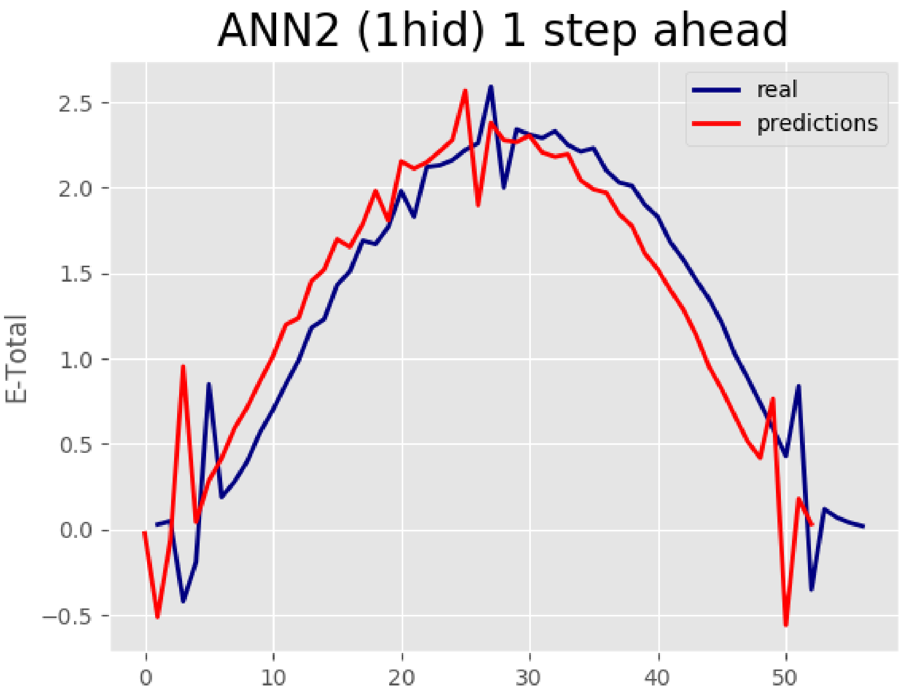
	ARIMA (1,1,0)	ARIMA (2,1,0)	ARIMA (3,1,0)
SMAPE (%)	11.5	16.87	21.11

As for the neural networks SMAPE metrics shows better results in ANN model of order 1 instead of order 2 or 3 that was also implemented. Table 11 and Figure 17 validates this conclusion presenting results from 1 step ahead, 3 inputs in the hidden layer and 1 output. The results of 2, 3 and 4 steps ahead are similar.

Table 11: SMAPE results of ANN models

	ANN - 1 neuron	ANN - 2 neurons	ANN – 3 neurons
SMAPE (%)	6.38	13.37	18.11


ANN (1hid) 1 step ahead
E-Total
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
0
10
20
30
40
50
real
predictions


ANN2 (1hid) 1 step ahead
E-Total
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
0
10
20
30
40
50
real
predictions

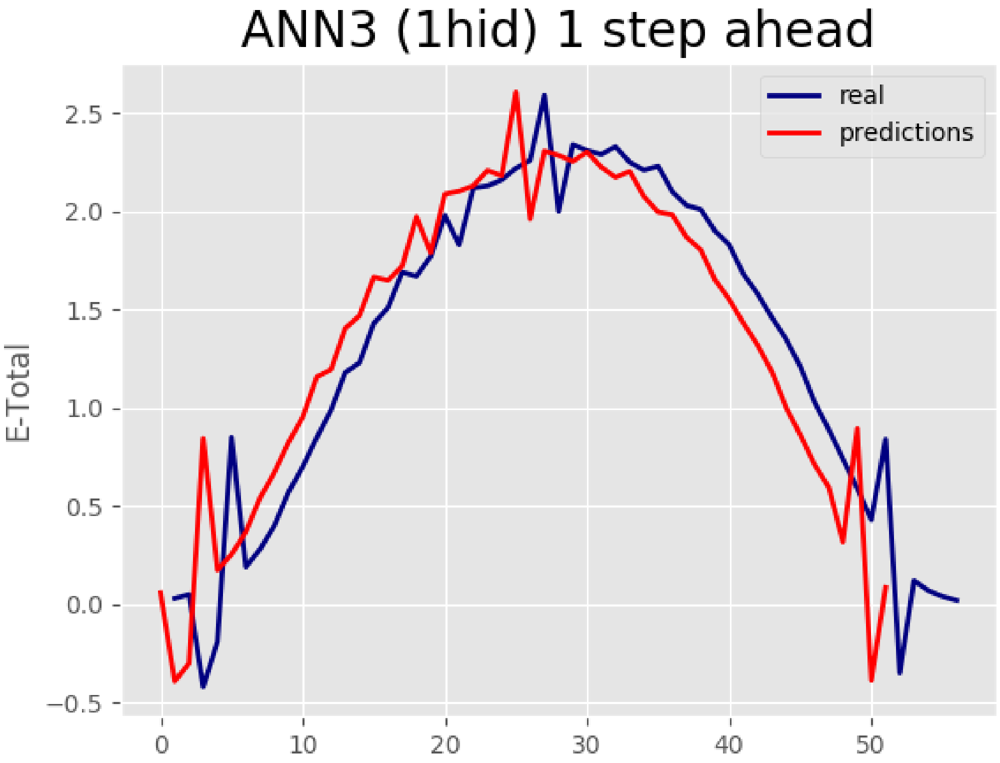


Figure 17: ANN models plot for one step ahead

In Figure 18, the fitting of the all the implemented models (ARIMA and ANNs, base and enhanced) on a test time series is presented. We can see that the forecasting results of all the implemented models fit very well on the real values. Finally, table 12 presents clearly the forecasting results of all the implemented models. As shown, ANN models have, in general, better forecasting accuracy compared to ARIMA. Also, we can see that the introduction of the meteorological data in the models as explanatory variables (i.e. enhanced models) decrease the forecasting accuracy in the case of the linear models (ARIMA), while it increases the forecasting accuracy in the case of non-linear models.

Table 12: Forecasting SMAPE Results of all implemented models

SMAPE (%)	Energy	En. + Irr.	En. + Amb. Temp.	En. + Panel's Temp.	Average
ARIMA	11.51	14.22	13.31	13.67	13.18
ANN	6.38	6.74	5.91	6.66	6.42

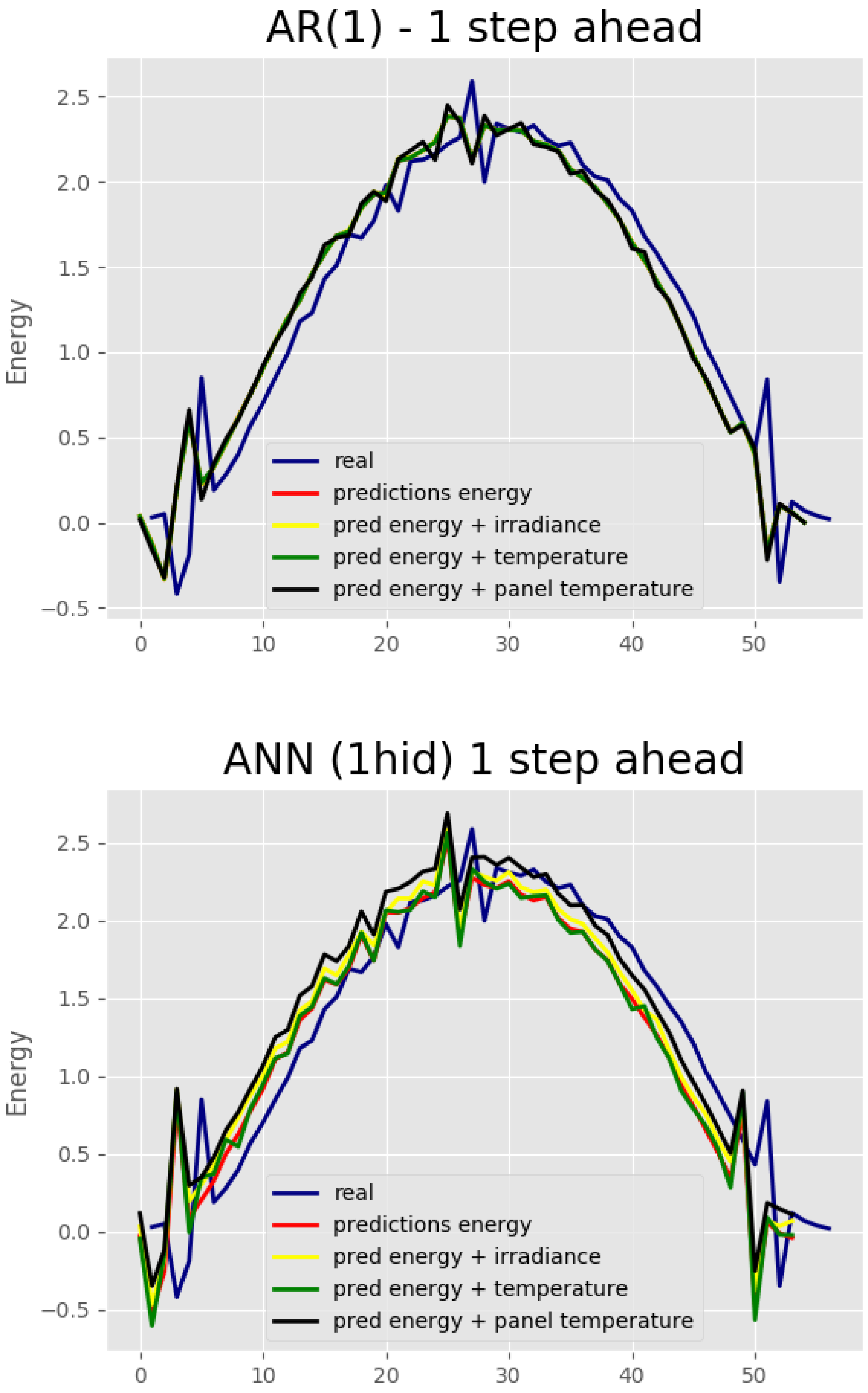


Figure 18: The ARIMA (above) and ANN (below), base and improved, models fitting a test time series.

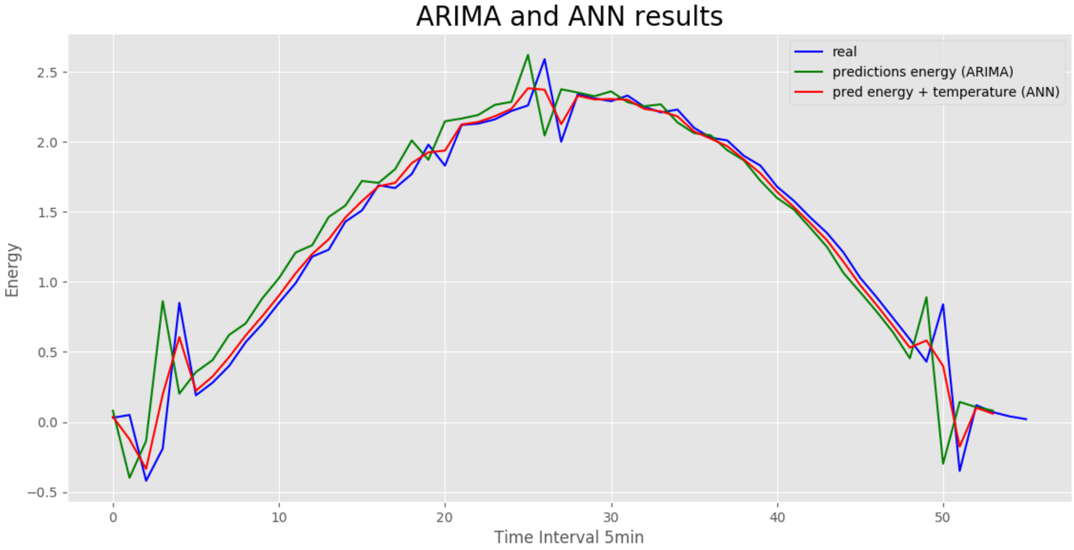
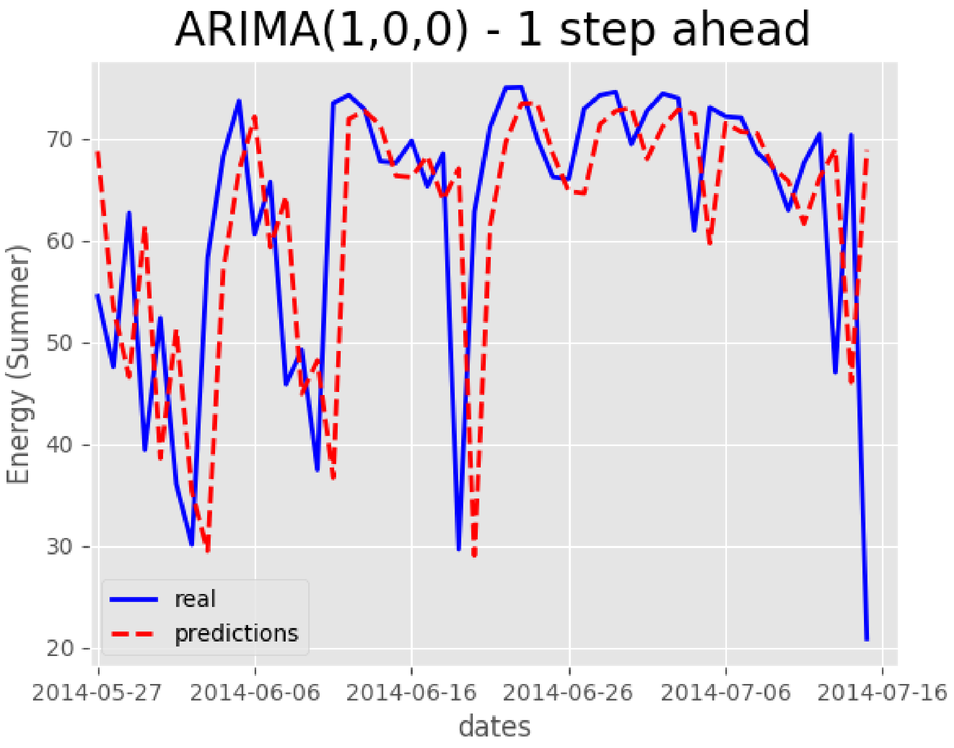
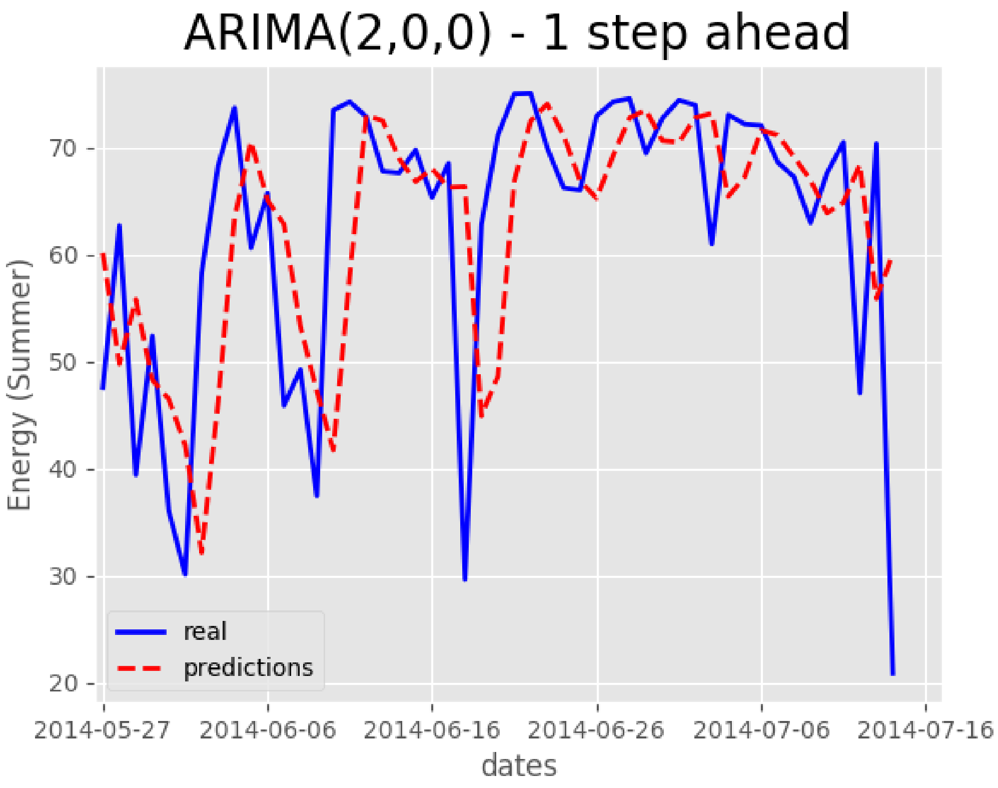


Figure 19: ARIMA and ANN best fitting

4.2. Long-Term Forecasting





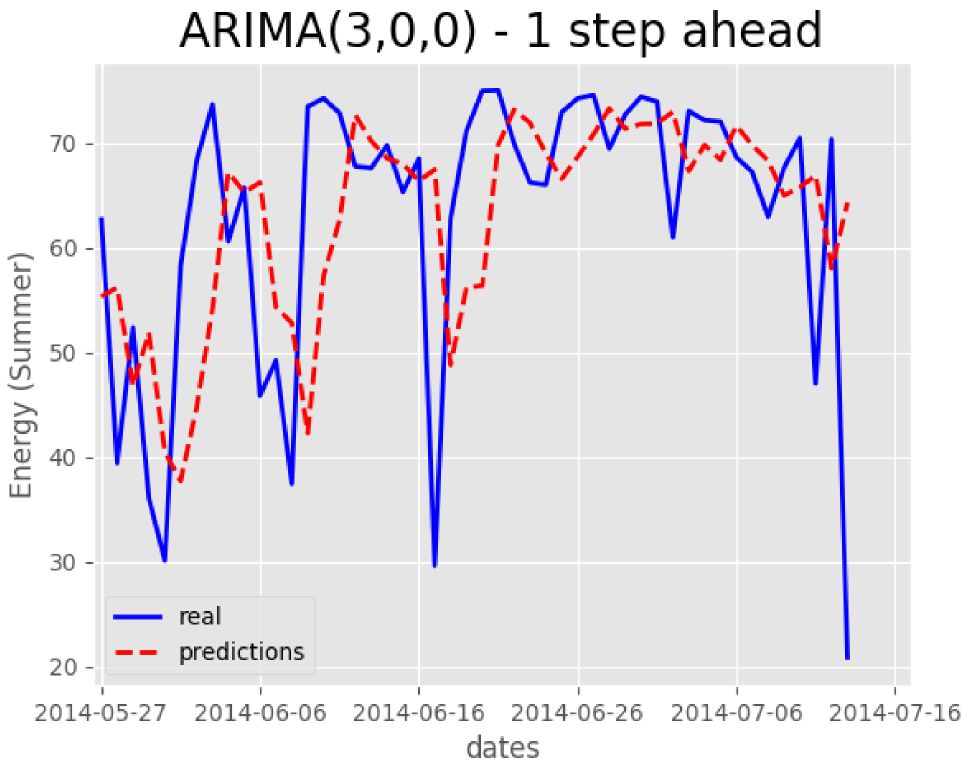
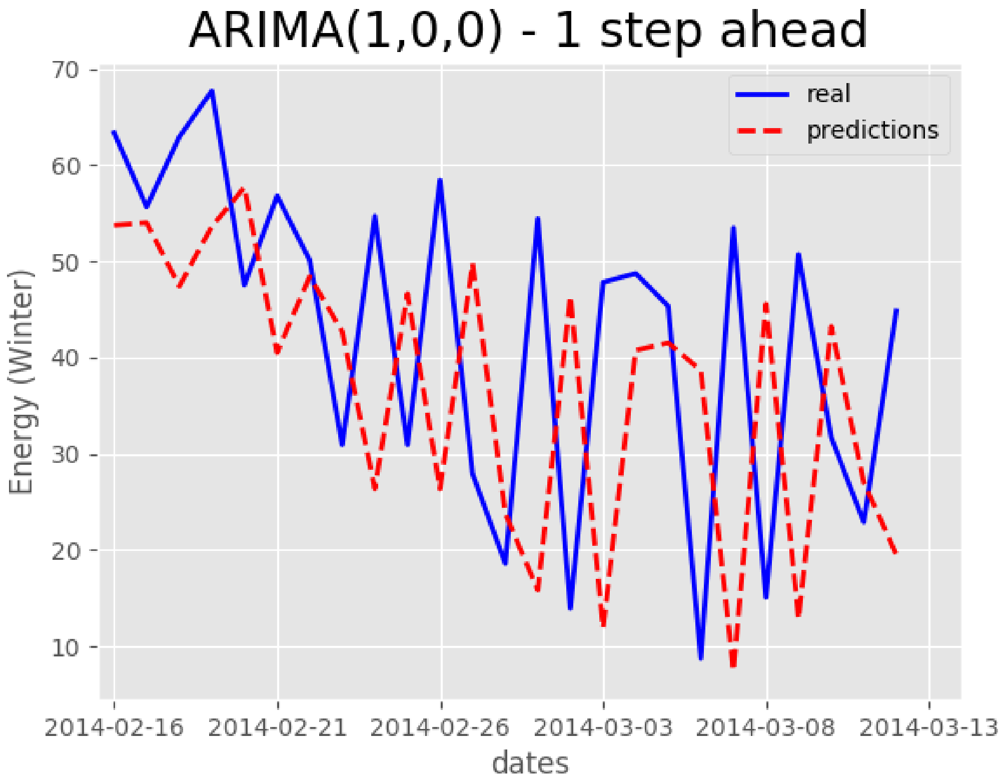


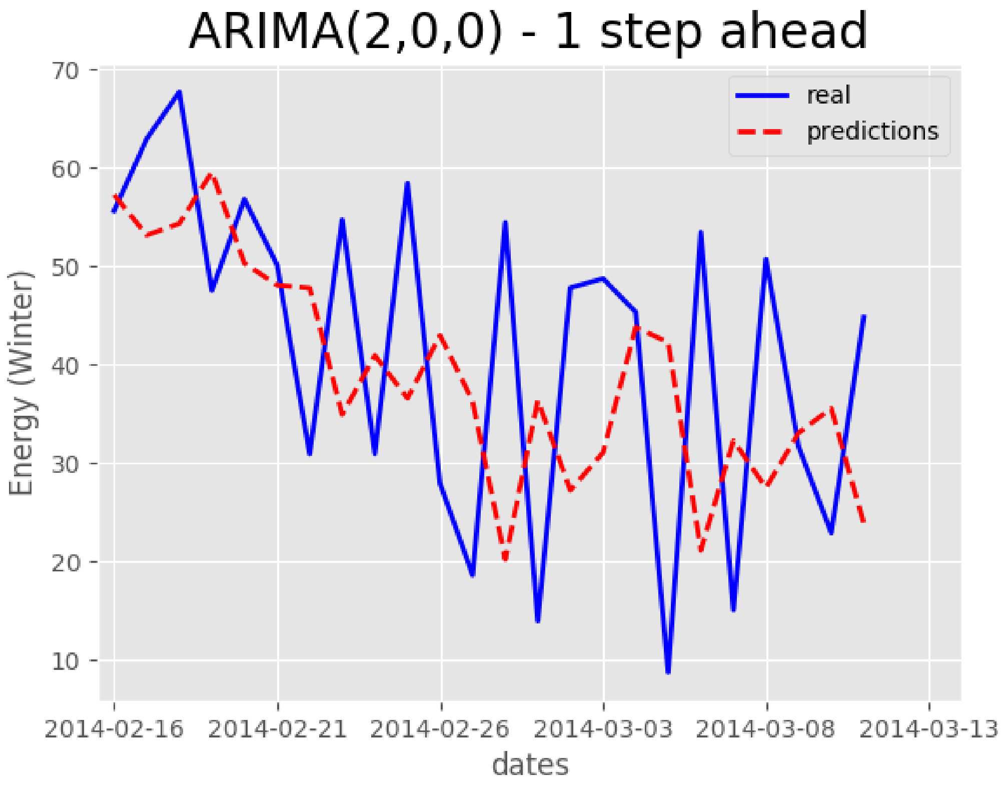
Figure 20: Plots of ARIMA models for 1 step ahead – 'Summer' period

In Figure 20 we can observe the ARIMA models fitting for 1 step ahead for the 'Summer' period and as we can see the ARIMA(2,0,0) fits better than its competitors. The results of 2, 3 and 4 steps ahead are similar. As a matter of fact, we confirmed it from the SMAPE results, as seen in Table 13.

Table 13: SMAPE results of ARIMA models for the 'Summer' period

Summer	**ARIMA (1,0,0)**	**ARIMA (2,0,0)**	**ARIMA (3,0,0)**
SMAPE (%)	9.42	8.87	9.14


ARIMA(1,0,0) - 1 step ahead
real
predictions
Energy (Winter)
70
60
50
40
30
20
10
2014-02-16
2014-02-21
2014-02-26
2014-03-03
2014-03-08
2014-03-13
dates


ARIMA(2,0,0) - 1 step ahead
real
predictions
Energy (Winter)
70
60
50
40
30
20
10
2014-02-16
2014-02-21
2014-02-26
2014-03-03
2014-03-08
2014-03-13
dates

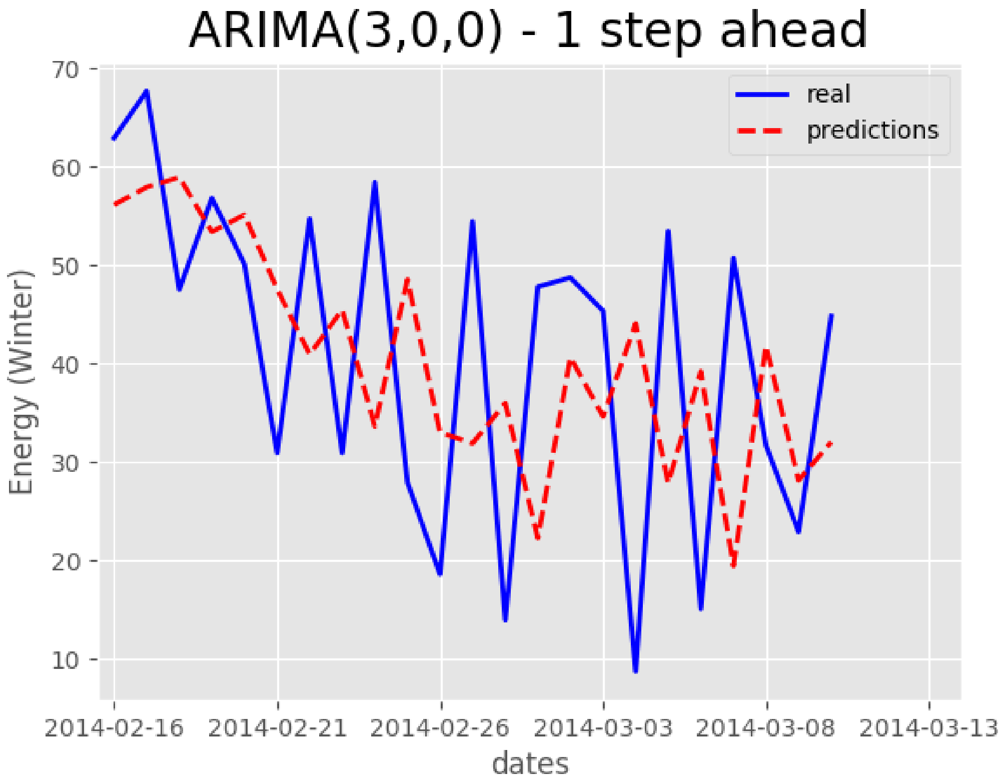
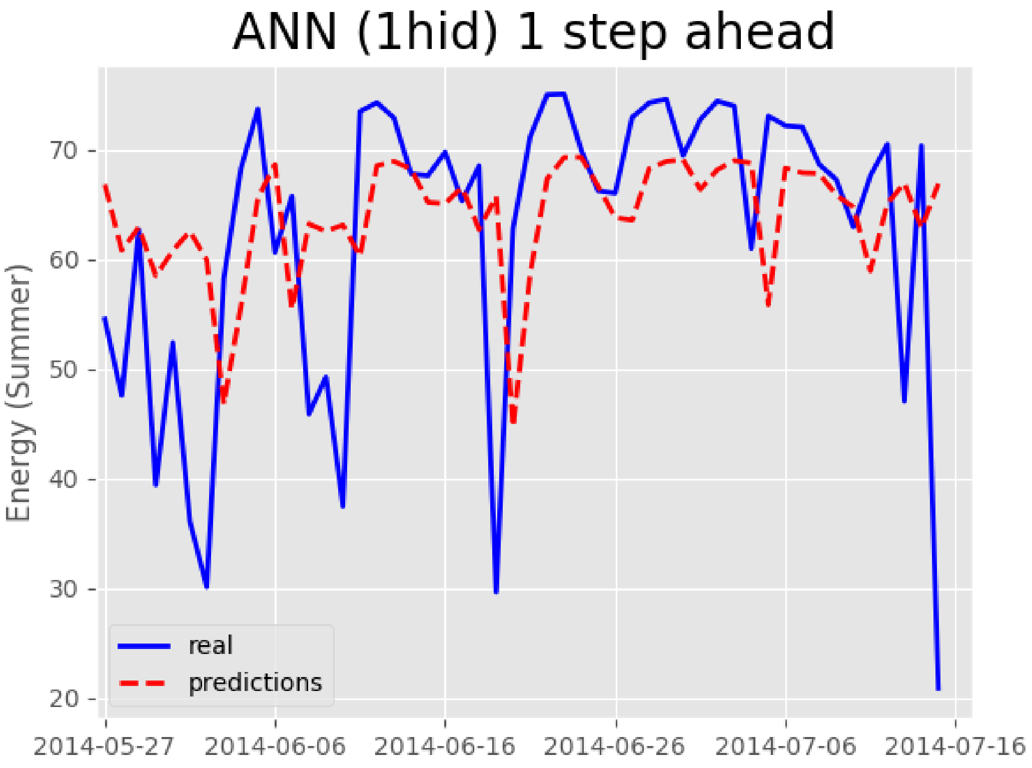


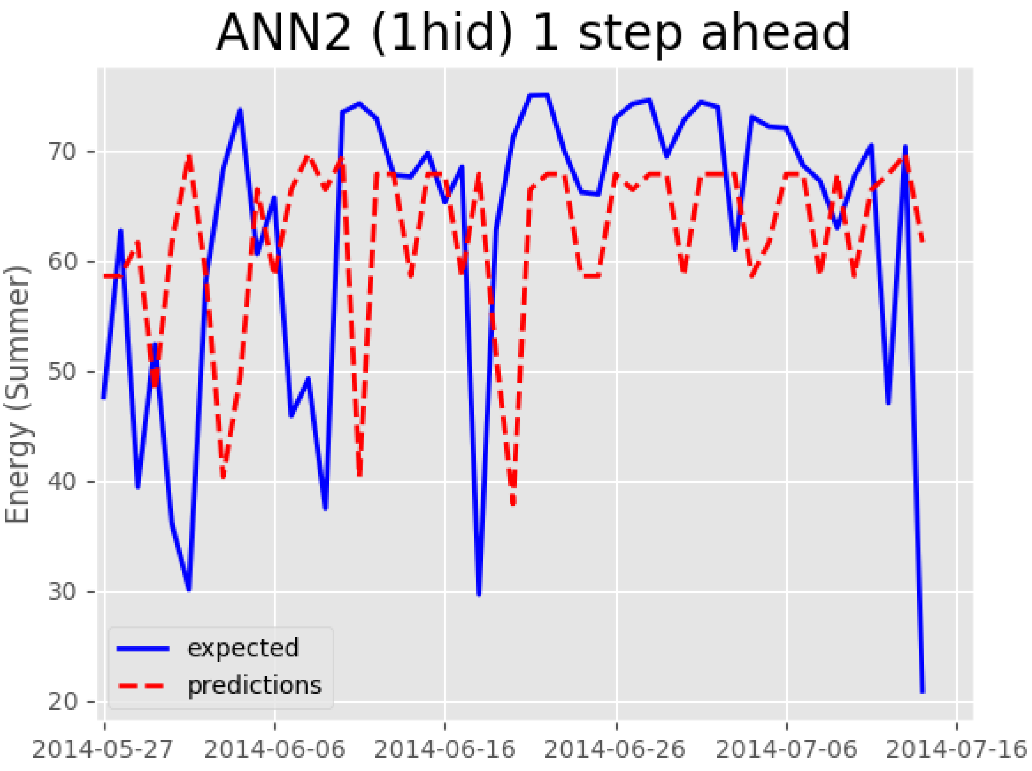
Figure 21: Plots of ARIMA models for 1 step ahead – 'Winter' period

In case of the 'Winter' period the ARIMA model that yield the best results is also the ARIMA (2,0,0) as in the 'Summer' period. From Figure 21 and Table 14 we verify the result.

Table 14: SMAPE results of ARIMA models for the 'Winter' period

Winter	**ARIMA (1,0,0)**	**ARIMA (2,0,0)**	**ARIMA (3,0,0)**
SMAPE (%)	28.26	22.71	22.85


ANN (1hid) 1 step ahead
Energy (Summer)
70
60
50
40
30
20
real
predictions
2014-05-27
2014-06-06
2014-06-16
2014-06-26
2014-07-06
2014-07-16


ANN2 (1hid) 1 step ahead
Energy (Summer)
70
60
50
40
30
20
expected
predictions
2014-05-27
2014-06-06
2014-06-16
2014-06-26
2014-07-06
2014-07-16

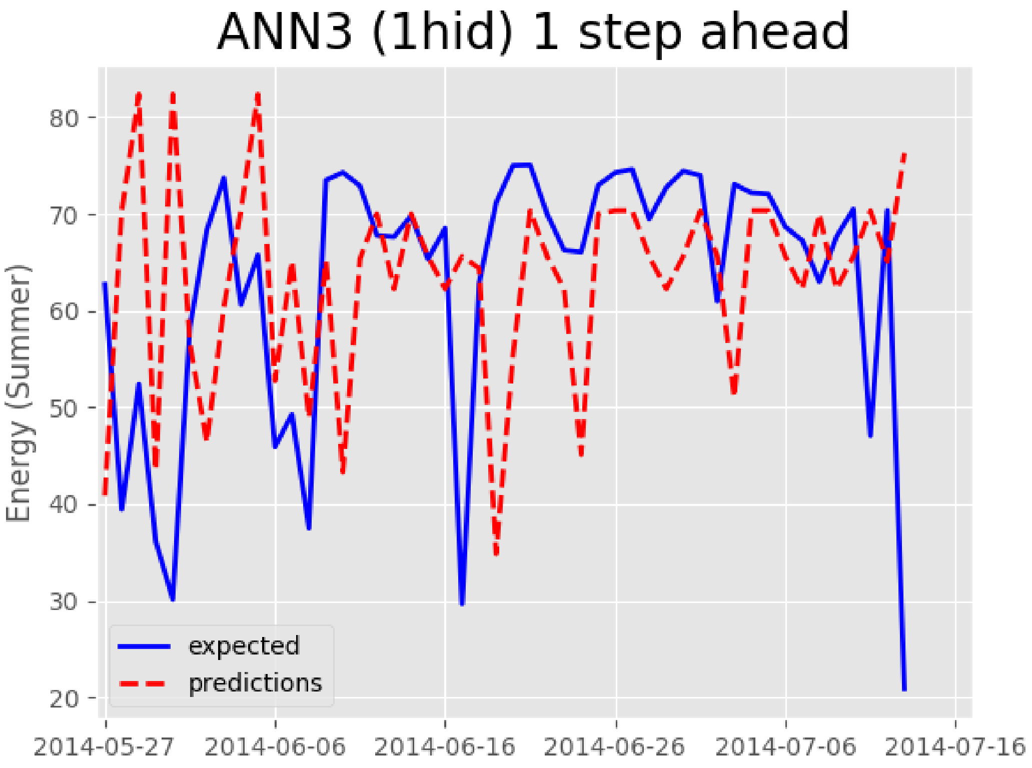
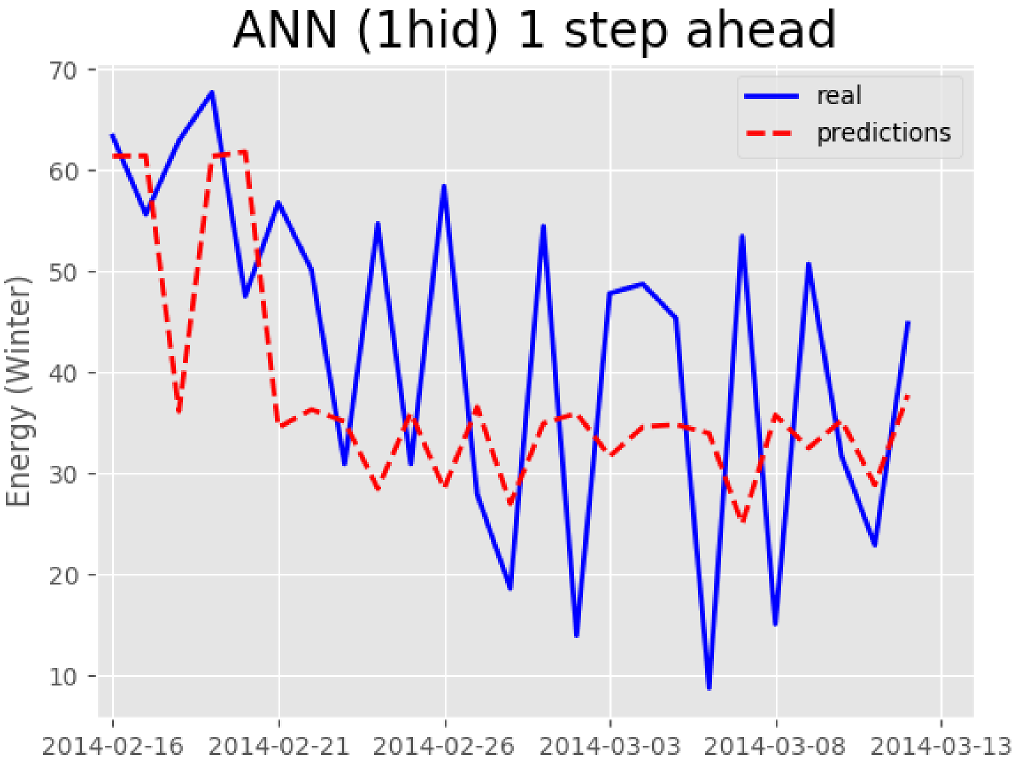


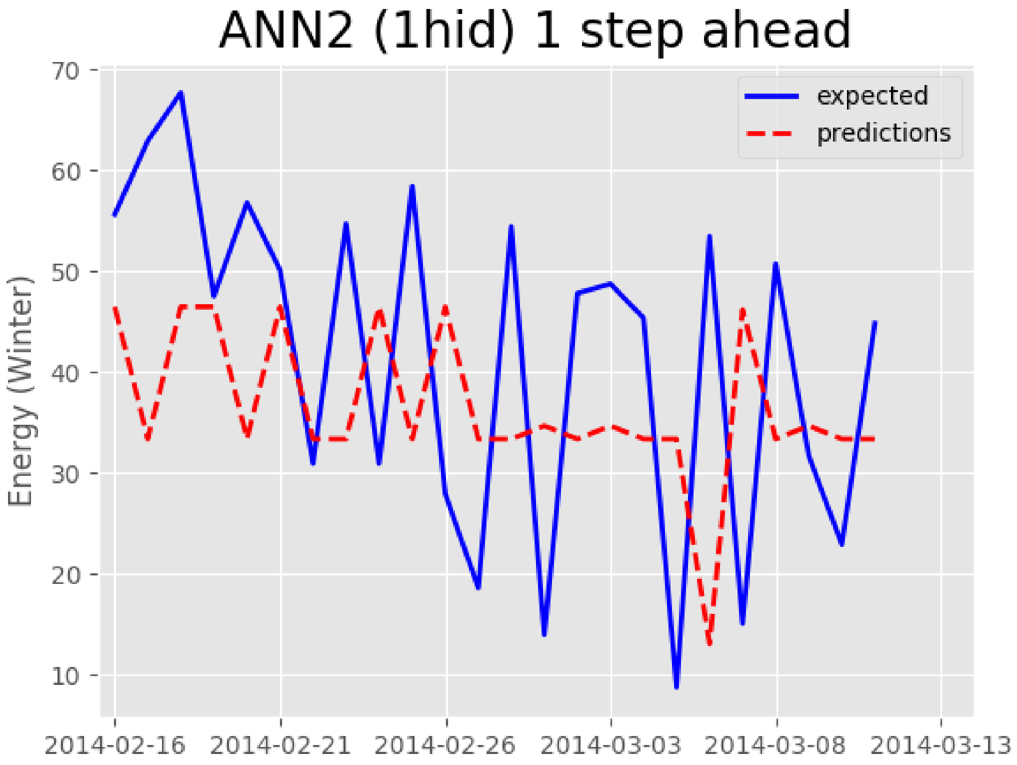
Figure 22: ANN plots of one step ahead – 'Summer' period

Figure 22 demonstrates the ANN's model fitting for the period of 'Summer'. As we can see from the plots and we can confirm from Table 15 the ANN with 1 neuron in the input layer gives the better results.

Table 15: SMAPE results of ANN models for the 'Summer' period

Summer	**ANN - 1 neuron**	**ANN - 2 neurons**	**ANN – 3 neurons**
SMAPE (%)	8.50	10.91	10.97


ANN (1hid) 1 step ahead
Energy (Winter)
70
60
50
40
30
20
10
2014-02-16
2014-02-21
2014-02-26
2014-03-03
2014-03-08
2014-03-13
real
predictions


ANN2 (1hid) 1 step ahead
Energy (Winter)
70
60
50
40
30
20
10
2014-02-16
2014-02-21
2014-02-26
2014-03-03
2014-03-08
2014-03-13
expected
predictions

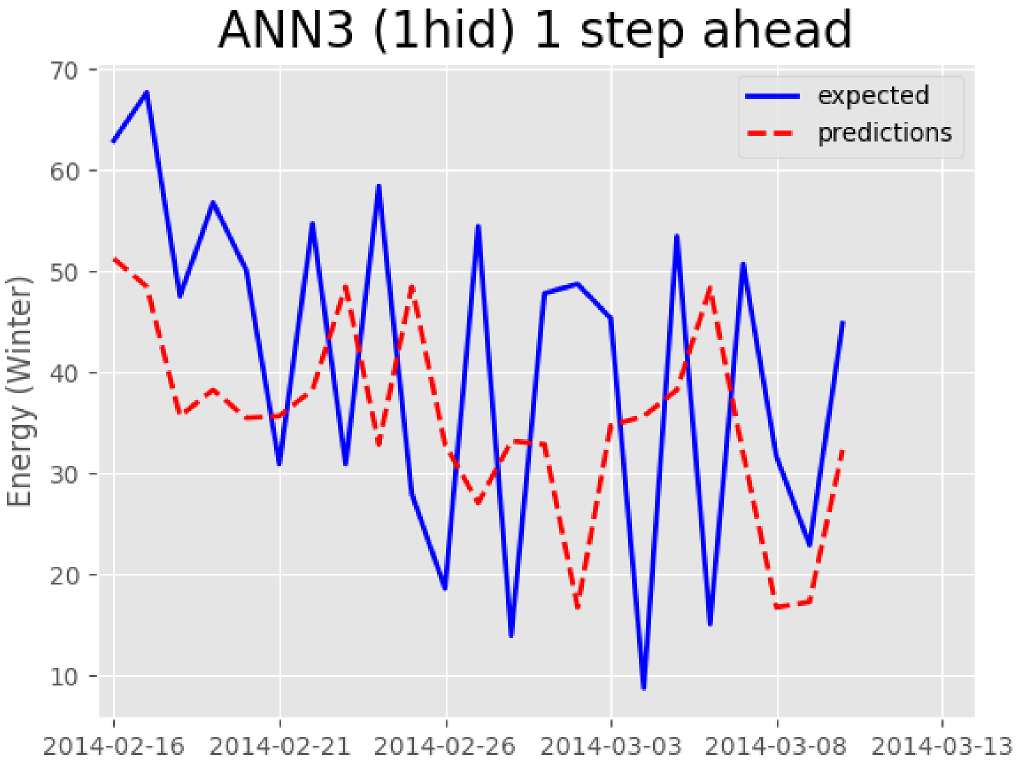


Figure 23: ANN plots for 1 step ahead – 'Winter' period

As for the 'Winter' period (Figure 23) the ANN with 1 neuron in the input layer fits better the original data. In Table 16, we can verify the result too.

Table 16: SMAPE results of ANN models for the 'Winter' period

Winter	**ANN - 1 neuron**	**ANN - 2 neurons**	**ANN – 3 neurons**
SMAPE (%)	20.12	23.44	25.03

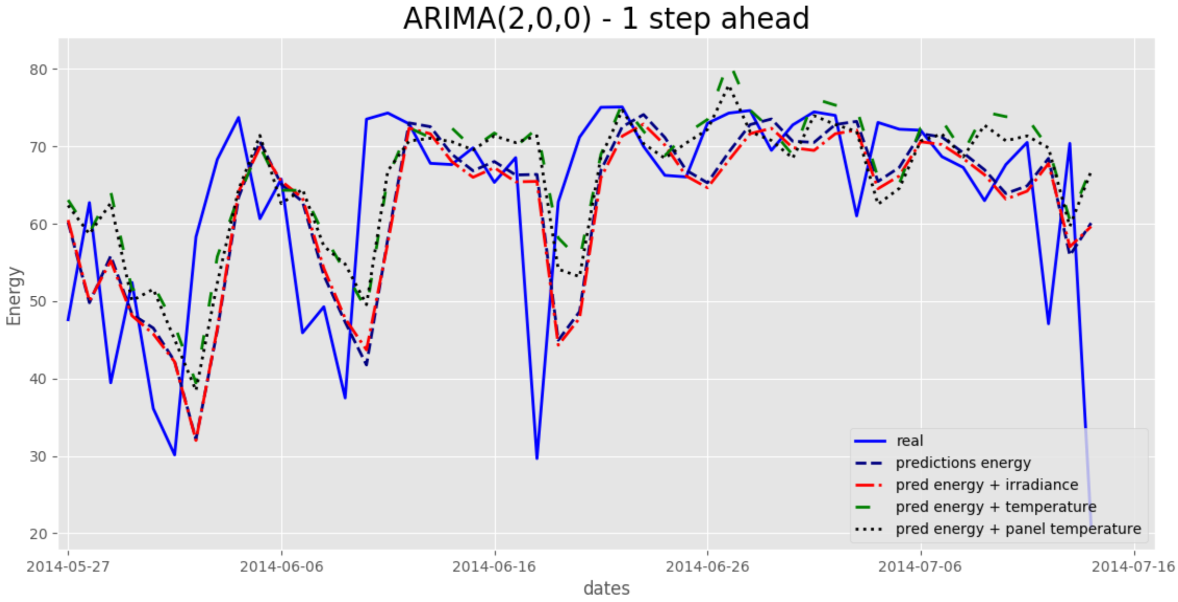


Figure 22: ARIMA fitting for all the implemented models –‘ Summer’ period

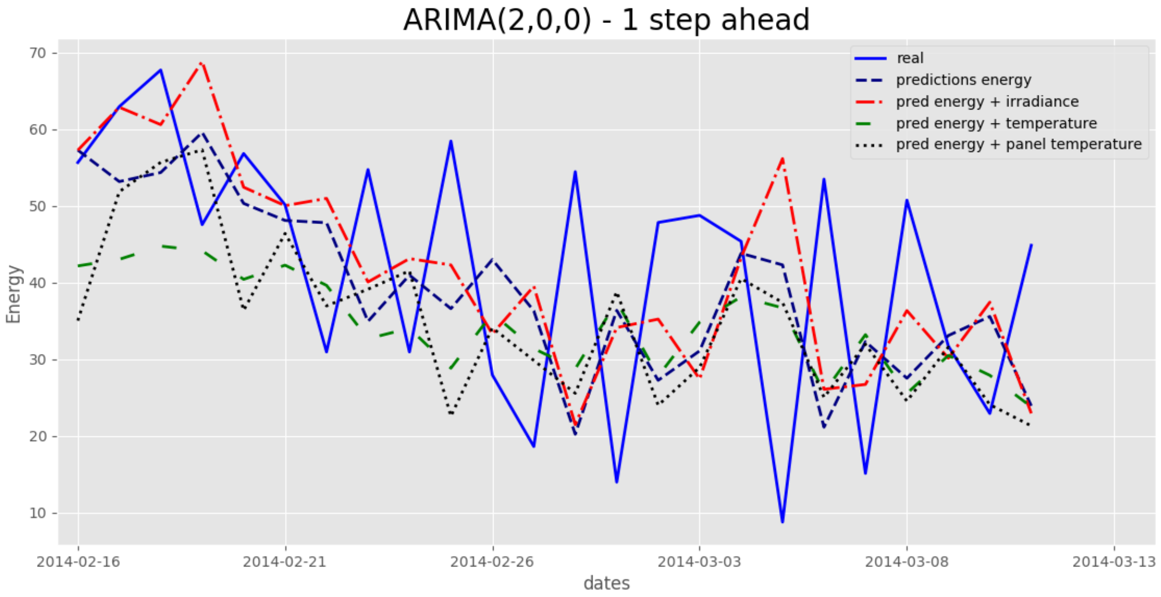


Figure 23: ARIMA fitting for all the implemented models – ‘Winter’ period

Table 17: SMAPE results for all the implemented models for the 'Summer' period

SMAPE (%)	Energy	En. + Irr.	En. + Amb. Temp.	En. + Panel's Temp.	Average
ARIMA	8.87	8.88	8.95	8.94	8.91
ANN	8.50	7.58	7.81	9.37	8.31

Table 18: SMAPE results for all the implemented models for the 'Winter' period

SMAPE (%)	Energy	En. + Irr.	En. + Amb. Temp.	En. + Panel's Temp.	Average
ARIMA	22.73	20.74	22.78	22.90	22.28
ANN	20.12	18.68	19.86	21.32	19.95

Figures 24 and 25 show the ARIMA model fitting for all the implemented models in both 'Summer' and 'Winter' periods. In general the period of 'Summer' is more accurate than the period of 'Winter'. The input of solar radiation, ambient temperature and panel's temperature as explanatory variables does not affect the accuracy in a tremendous level. There is a slight decrease in terms of accuracy.

The following Figures 26 and 27, present the ANN model fitting for all the implemented models in the periods 'Summer' and 'Winter'. We observed that the 'Summer' period has better accuracy contrary to the 'Winter' period as in the ARIMA model fitting. The fusion of meteorological values does affect the accuracy.

In general ANN models yielded better results contrary to ARIMA models, especially in the 'Winter' period. In the 'Summer' period the ANN models surpass the ARIMA models too, but with a slight difference. All these conclusions were verified from Table 17 and Table 18. The best results are represented in Figures 28 and 29.

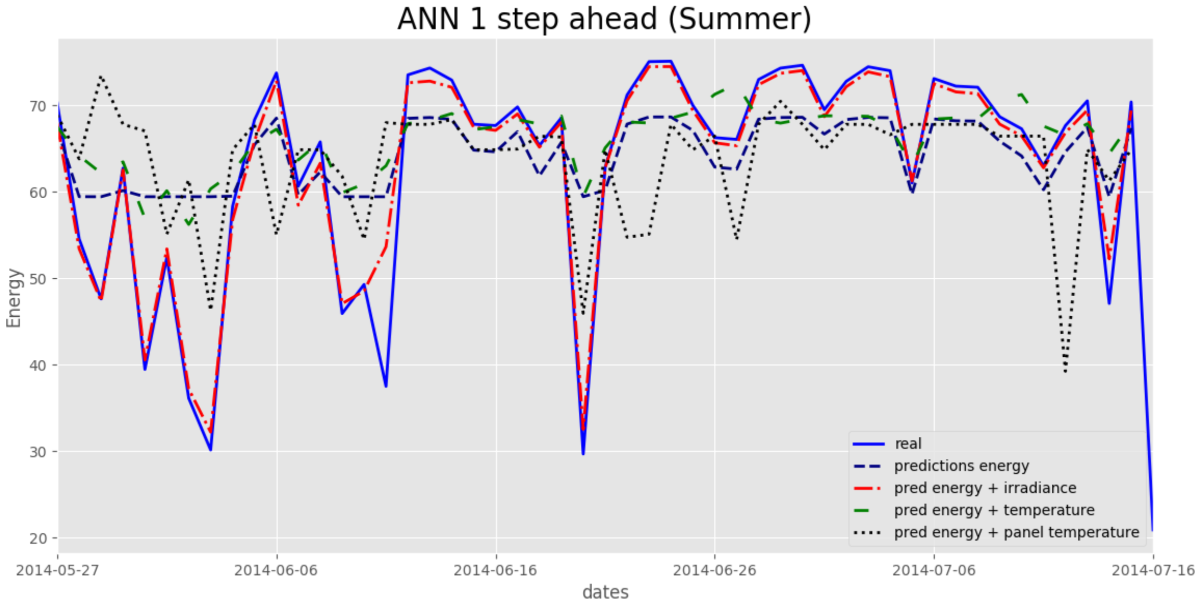


Figure 24: ANN fitting for all the implemented models of period 'Summer'

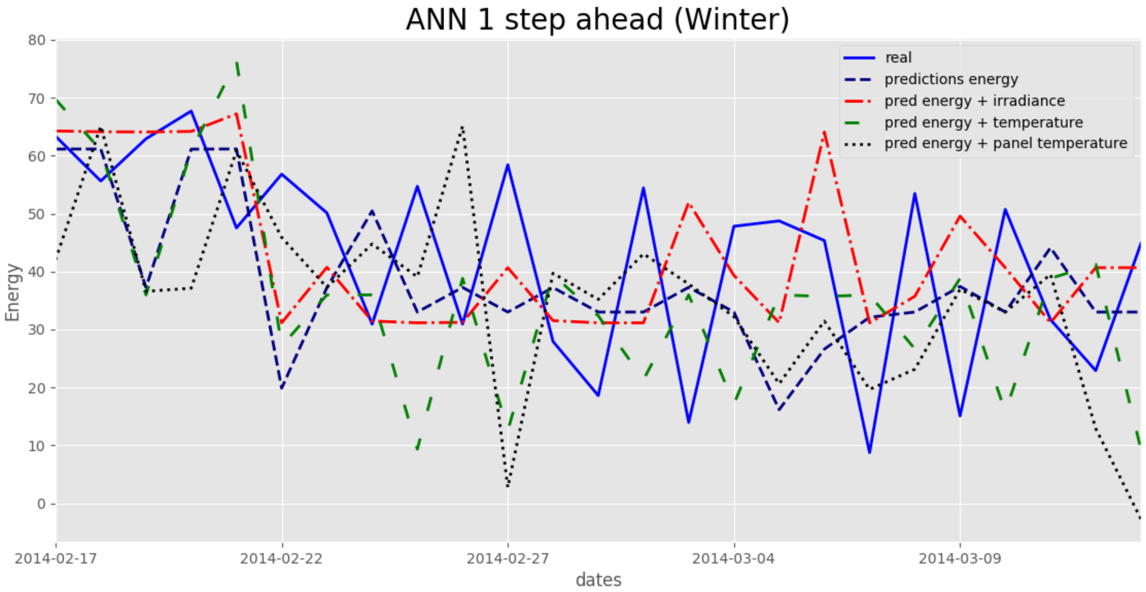


Figure 25: ANN fitting for all the implemented models of period 'Winter'

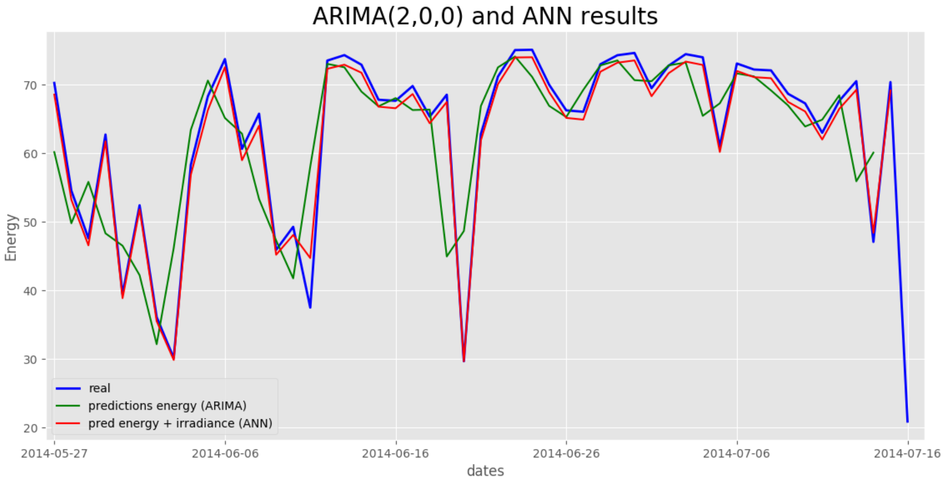


Figure 26: ARIMA and ANN best fitting of "Summer" period

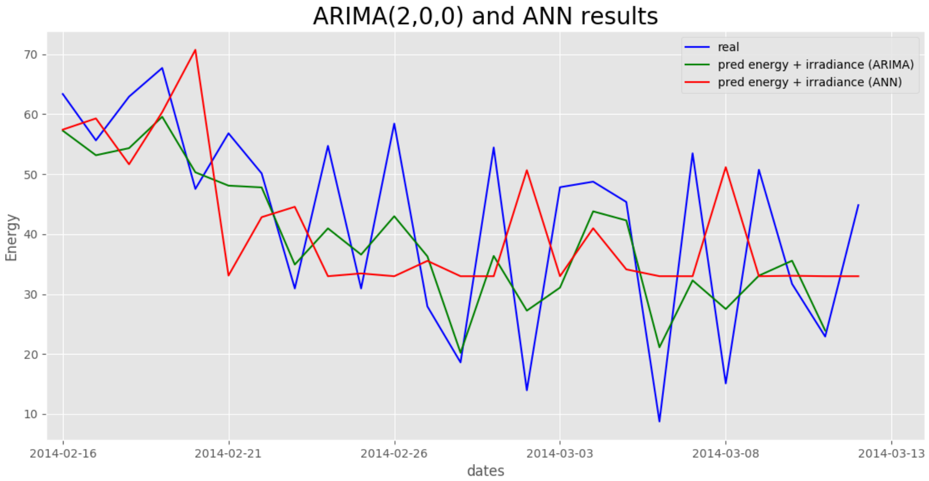


Figure 27: ARIMA and ANN best fitting of "Winter" period

CONCLUSIONS

A set of different models for forecasting power output in PV systems were implemented and evaluated. All steps of time series modelling process were implemented and presented in detail, along with the corresponding steps for building non-linear models. Moreover, the models were enhanced with meteorological explanatory variables.

Preliminary results, for experiments conducted on real data from a photovoltaic park in Crete, Greece, indicated that the ANN non-linear models outperform the linear ARIMA models in terms of forecasting accuracy. Additionally, the fusion of meteorological data increases the forecasting accuracy of the ANN models, while decreasing the accuracy of ARIMA models. The proposed ANN topology, which performs best is an ANN with 1 neuron in the input layer, 3 in the hidden and 1 in the output layer. Furthermore, a long-term forecast was conducted from the same data and new models for forecasting power output in PV system were implemented and evaluated. Also in this analysis the ANN models yielded the best results, contrary to ARIMA. In case of long-term forecasting analysis there is not a specific proposed topology of the ANN model due to the fact that, for every implemented model, base and improved, there is a different topology. Future directions of this work include, the experimentation with more data from different PV systems, the fusion of new data variables in the implemented models and the implementation of a hybrid model that will exploit the best characteristics of the ARIMA and ANN models. The examination of the possibility of arising a problem of overfitting, is also a recommendation of future work.

PUBLICATIONS

Part of this master thesis and specifically the short-term analysis has been accepted for presentation in the International work-conference on Time Series (ITISE) 2017 which was held in Granada, Spain in September 18th-20th, 2017. The paper has the title ‘Forecasting Power Output of Photovoltaic Systems Using Linear, Non-Linear and Enhanced Models’ and it is included in the ITISE-2017 proceedings.

BIBLIOGRAPHY

[1] Ahmia, O., Farah, N., & Mokhtar, B. (2015). Parallel seasonal approach for electrical load forecasting., (July). https://doi.org/10.13140/RG.2.1.1504.3680

[2] Basu, S., & Meckesheimer, M. (2007). Automatic outlier detection for time series: An application to sensor data. *Knowledge and Information Systems*, *11*(2), 137–154. https://doi.org/10.1007/s10115-006-0026-6

[3] Chen, C., Duan, S., Cai, T., & Liu, B. (2011). Online 24-h solar power forecasting based on weather type classification using artificial neural network. *Solar Energy*, *85*(11), 2856–2870. https://doi.org/10.1016/j.solener.2011.08.027

[4] Gandelli, A., Grimaccia, F., Leva, S., Mussetta, M., & Ogliari, E. (2014). Hybrid model analysis and validation for PV energy production forecasting. In *Proceedings of the International Joint Conference on Neural Networks* (pp. 1957–1962). https://doi.org/10.1109/IJCNN.2014.6889786

[5] Hamid Oudjana, S., Hellal, a., & Hadj Mahamed, I. (2012). Short term photovoltaic power generation forecasting using neural network. *2012 11th International Conference on Environment and Electrical Engineering*, pp. 706–711. https://doi.org/10.1109/EEEIC.2012.6221469

[6] Jiahao, K., Jun, L., Qifan, L., Wanliang, F., Zhenhuan, C., Linlin, L., & Tieying, G. (2013). Photovoltaic power forecasting based on artificial neural network and meteorological data. In *TENCON 2013 - 2013 IEEE Region 10 Conference (31194)* (pp. 1–4). https://doi.org/10.1109/TENCON.2013.6718512

[7] Kou, J., Liu, J., Li, Q., Fang, W., Chen, Z., Liu, L., & Guan, T. (2013). Photovoltaic power forecasting based on artificial neural network and meteorological data. *2013 IEEE International Conference of IEEE Region 10 (TENCON 2013)*, pp. 1–4. https://doi.org/10.1109/TENCON.2013.6718512

[8] Lee, W., & Wang, P. (2011). Forecasting Power Output of Photovoltaic Systems Based on Weather Classification and Support Vector Machines. *Industry Applications Society Annual Meeting*, *48*(c), pp. 1–6.

[9] Leva, S., Dolara, A., Grimaccia, F., Mussetta, M., & Ogliari, E. (2017). Analysis and validation of 24 hours ahead neural network forecasting of photovoltaic output power. *Mathematics and Computers in Simulation*, *131*, pp. 88–100. https://doi.org/10.1016/j.matcom.2015.05.010

[10] Lo Brano, V., Ciulla, G., & Di Falco, M. (2014). Artificial Neural Networks to Predict the Power Output of a PV Panel. *International Journal of Photoenergy*, *2014*, 12. https://doi.org/10.1155/2014/193083

[11] Makridiakis, S., & Hibon, M. (1997). ARMA Models and the Box-Jenkins Methodology. *Journal of Forecasting*, *16*(3), pp. 147–163. https://doi.org/10.1002/(SICI)1099-131X(199705)16:3<147::AID-FOR652>3.0.CO;2-X

[12] Malvoni, M., De Giorgi, M. G., & Congedo, P. M. (2014). Photovoltaic power forecasting using statistical methods: impact of weather data. *IET Science, Measurement & Technology*, *8*(3), pp. 90–97. https://doi.org/10.1049/iet-smt.2013.0135

[13] Mellit, A., Sağlam, S., & Kalogirou, S. A. (2013). Artificial neural network-based model for estimating the produced power ofaphotovoltaic module. *Renewable Energy*, *60*, pp. 71–78. https://doi.org/10.1016/j.renene.2013.04.011

[14] Pedro, H. T. C., & Coimbra, C. F. M. (2012). Assessment of forecasting techniques for solar power production with no exogenous inputs. *Solar Energy*, *86*(7), pp. 2017–2028. https://doi.org/10.1016/j.solener.2012.04.004

[15] Rai, A. K., Kaushika, N. D., Singh, B., & Agarwal, N. (2011). Simulation model of ANN based maximum power point tracking controller for solar PV system. *Solar Energy Materials and Solar Cells*, *95*(2), pp. 773–778. https://doi.org/10.1016/j.solmat.2010.10.022

[16] Shi, J., Lee, W. J., Liu, Y., Yang, Y., & Wang, P. (2012). Forecasting power output of photovoltaic systems based on weather classification and support vector machines. In *IEEE Transactions on Industry Applications* (Vol. 48, pp. 1064–1069). https://doi.org/10.1109/TIA.2012.2190816

[17] Teo, T. T., Logenthiran, T., & Woo, W. L. (2016). Forecasting of photovoltaic power using extreme learning machine. In *Proceedings of the 2015 IEEE Innovative Smart Grid Technologies - Asia, ISGT ASIA 2015*. https://doi.org/10.1109/ISGT-Asia.2015.7387113

[18] Xiyun Yang, Feifei Jiang, & Huan Liu. (2013). Short-Term Solar Radiation Prediction based on SVM with Similar Data. *2nd IET Renewable Power Generation Conference (RPG 2013),* (2), 1.11-1.11. https://doi.org/10.1049/cp.2013.1735

[19] Yang, X., Jiang, F., & Liu, H. (2013). Short-Term Power Prediction of Photovoltaic Plant Based on SVM with Similar Data and Wavelet Analysis, (5), 81–85.

[20] Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, *50*, pp. 159–175. https://doi.org/10.1016/S0925-2312(01)00702-0