Tampereen yliopisto

Väinö Anttalainen

AURINKOENERGIAJÄRJESTELMÄN TEHON MALLINTAMINEN

Fysikaalisten ja koneoppimispohjaisten tehomallien vertailu

Diplomityö
Tekniikan ja luonnontieteiden tiedekunta
Tarkastajat: Prof. Tarmo Lipping (TUNI)
TkT Sami Jouttijärvi (UTU)
Prof. Juho Kanniainen (TUNI)
Maaliskuu 2025

TIIVISTELMÄ

Väinö Anttalainen: Aurinkoenergiajärjestelmän tehon mallintaminen
Diplomityö
Tampereen yliopisto
Teknis-luonnontieteellinen DI-ohjelma
Maaliskuu 2025

Joissain maissa aurinkoenergia on jo valtavirtaa, kun taas jossain maissa, kuten Suomessa, aurinkoenergia on vasta nousemassa aurinkovoimateknologian kehittyessä ja halventuessa. Jotta aurinkovoimalat toimisivat odotetusti, niiden tehon tuotantoa on tärkeää monitoroida vertaamalla todellista tehon tuotantoa mallinnettuun tehon tuotantoon. Koska yleisesti käytettyjen fysikaalisten tehomallien on näytetty antavan Suomen oloissa virheellisiä tuloksia esimerkiksi lumipeitteen ja alhaisen auringon säteilyn vuoksi erityisesti talvisin, työssä kokeillaan myös monimutkaisempia koneoppimispohjaisia tehomalleja. Aineistoa on viidestä aurinkoenergiajärjestelmästä neljästä eri sijainnista (Helsingistä, Turusta, Kuopiosta ja Sodankylästä).

Tämän työn tavoitteena on tarkastella tehomallien virhettä hetkittäisessä tehon tuotannossa sekä energian tuotannon virhettä kuukausi- ja vuositasolla. Lisäksi työssä pyritään selvittämään, pystytäänkö koneoppimismalleilla mallintamaan tehoa tarkemmin kuin fysikaalisilla malleilla. Työssä tarkastellaan kolmea fysikaalista mallia ja kahta koneoppimismallia.

Jotkin fysikaaliset mallit voidaan optimoida tiettyyn aurinkovoimajärjestelmään sopivaksi mikäli aiempaa aineistoa järjestelmästä on saatavilla. Tällaisia malleja kutsutaan tässä työssä täsmäkoulututetuiksi malleiksi. Jos aiempaa aineistoa ei ole saatavilla, on tehon mallintamiseen käytettävä yleisempiä fysikaalisia malleja, joita ei ole optimoitu kyseiseen järjestelmään. Tällaisia malleja kutsutaan tässä työssä yleisiksi malleiksi.

Tulosten perusteella yleiset fysikaaliset mallit näyttäisivät mallintavan tehon systemaattisesti liian suureksi, mikä kasvattaa virheitä myös energian tuotannon mallintamisessa. Käytettyjen virhetermien mukaan yleiset koneoppimismallit mallintavat tehoa tarkemmin kuin yleiset fysikaaliset mallit. Yleisillä koneoppimismalleilla ei ole samanlaista taipumusta mallintaa tehoa aina liian suureksi, mutta koneoppimismallit näyttäisivät mallintavan alhaiset tehon arvot liian suuriksi ja suuret tehon arvot liian alhaisiksi. Koska osa virheistä kumoutuu, on kuukausittaiset ja vuosittaiset energian tuotannon mallintamisen virheet fysikaalisia malleja pienempiä.

Täsmäkoulutus näyttäisi parantavan fysikaalisten ja koneoppimismallien tarkkuuksia ja vähentävän virheitä sekä tehon että energian tuotannon mallintamisessa. Täsmäkoulutuksen jälkeen toinen koneoppimismalleista antoi kaikista malleista pienimmät virhetermit, mutta tämän mallin kuukausittaiset ja vuosittaiset energian tuotannon mallinnuksen virheet yllättävästi kasvoivat pienentymisen sijaan.

Vaikka koneoppimismallit näyttäisivät virhetermien ja energian tuotannon virheiden perusteella toimivan fysikaalisia malleja paremmin, ei mallien tarkkuuksien ero vaikuta olevan kovin suuri. Suurimmat erot fysikaalisten ja koneoppimismallien välillä näyttäisivät olevan mallien monimutkaisuudessa ja siinä, millainen systemaattinen taipumus yleisillä malleilla on tehon mallintamisessa. Koska koneoppimismallit tarvitsevat havaintoja useammasta muuttujasta toimiakseen, voi niiden käyttö olla hankalampaa etenkin pienemmillä aurinkoenergiajärjestelmillä, joissa on vähemmän mittalaitteita. Koneoppimismallien käyttö voi olla kuitenkin hyödyllistä esimerkiksi kaupallisilla aurinkovoimajärjestelmillä, joissa halutaan monitoroida tehoa mahdollisimman tarkasti tehomallin monimutkaisuudesta huolimatta.

Avainsanat: aurinkoenergia, PV, tehon mallintaminen, fysikaalinen malli, koneoppiminen

ABSTRACT

Väinö Anttalainen: Modeling the power of the solar energy system
Master's thesis
Tampere University
Master's Programme in Science and Engineering
March 2025

In some countries solar energy is already mainstream while in others, like Finland, it is gaining popularity due to advancements and decreasing costs in solar energy technologies. To ensure that the solar power plants operate as expected, it is essential to monitor them by comparing the expected power production to the actual power production. Because commonly used power models have been shown to produce erroneous results in Finland, more complex machine learning models are also implemented in this thesis. The data used is from five different power plants across four locations: Helsinki, Turku, Kuopio, and Sodankylä.

The aim of this thesis is to study the errors that the models produce when modelling the instantaneous power productions, as well as the monthly and yearly energy yields. In addition, machine learning models are compared to physical models to determine if the added complexities of machine learning result in more accurate estimates. The comparison includes three physical models and two machine learning models. Some physical models can be optimized for a specific power system if data is available from that system. These models are referred to as fitted models. If previous data is not available, the models can't be optimized to the system. These models are referred as general models.

Results indicate that the general physical models systematically overestimate produced power, leading to increased errors in estimated energy yields. Based on the used error terms, general machine learning models predict power more accurately than general physical models. Machine learning models tend to overestimate low power values and underestimate high power values, causing some errors to cancel each other out and resulting in lower errors in estimated energy yields compared to physical models.

Fitted models seem to be more accurate than general models, which is expected. One machine learning model outperformed all fitted models based on the error terms used. However, errors in monthly and yearly energy yield estimations increased unexpectedly.

Although machine learning models appear to model power more accurately than physical models, whether fitted or general, the differences are small. The biggest differences between physical and machine learning models seem to be the complexities of the models and the systematic tendencies of general models. Because machine learning models require observations from more variables than physical models, their use can be more challenging, especially for smaller power plants with fewer measuring tools. However, the slight accuracy gains can be beneficial for larger commercial power plants despite the added model complexity.

Keywords: solar energy, photovoltaics, power modelling, physical model, machine learning

TEKOÄLYN KÄYTTÖ

Opinnäytteessäni käytetyt tekoälytyökalut ja niiden käyttötarkoitukset on kuvailtu alla:

Työkalun nimi: Microsoft Copilot TUNI-tunnuksella kirjautuneena.

Työkalua on käytetty

- apuna työssä käytettujen Python-koodien kirjoittamiseen,
- apuna luvun 5 taulukoiden muotoilussa,
- englanninkielisen tiivistelmän kielen tarkistamisessa.

Olen tietoinen siitä, että olen täysin vastuussa koko opinnäytteeni sisällöstä, mukaan lukien tekoälyllä tuotetut osat, ja hyväksyn vastuun mahdollisista eettisten ohjeiden rikkomuksista.

ALKUSANAT

Työ on tehty Turun yliopiston materiaalitekniikan yksikössä Solar Energy Materials and Systems (SEMS) -tutkimusryhmässä, jota johtaa professori Kati Miettunen. Työ liittyy RealSolar-hankkeeseen, jota rahoittaa Suomen Akatemian Strategisen tutkimuksen neuvosto.

Kiitän ohjaajiani Tarmo Lippingiä Tampereen yliopistosta ja Sami Jouttijärveä Turun yliopistosta ohjauksesta, tuesta ja ideoista. Vaikka Kati Miettunen ja Lauri Karttunen eivät virallisesti olleetkaan ohjaajiani, iso kiitos kuuluu myös heille tämän työn mahdollistamisesta. Olen myös kiitollinen muista kollegoista Turun yliopistossa, joihin olen työn kirjoittamisen myötä tutustunut. Teidän takia on ollut ilo tulla töihin joka päivä. Kiitän myös Ilmatieteen laitosta ja Turun ammattikorkeakoulua työssä käytetyistä aineistoista sekä CSC:tä, jonka Puhti-supertietokonetta olen käyttänyt koneoppimismallien kouluttamiseen. Eikä tätäkään työtä olisi syntynyt ilman perhettäni. Kiitos teille avusta ja tuesta sekä ilojen ja surujen jakamisesta.

Tampereella, 13. maaliskuuta 2025

Väinö Anttalainen

SISÄLLYSLUETTELO

LYHENTEET JA MERKINNÄT

C	lämpötilakorjaustermi
E	PV-järjestelmän energia
F_1, F_2	Perezin diffuusiomalliin liittyvät kirkkausmallit
G_b	paneelille osuvan säteilyn suora komponentti
G_d	paneelille osuvan säteilyn diffusoitunut komponentti
G_g	paneelille osuvan säteilyn maasta heijastuva komponentti
G_{DHI}	diffusoituneen horisontaalisäteilyn arvo
G_{DNI}	suoran normaalisäteilyn arvo
G_{GHI}	globaalin horisontaalisäteilyn arvo
G_{POA}	paneelin pinnan mukaisen säteilyn arvo
H	gradienttiboostauksessa käytettävä heikko oppija (*weak learner*)
I	sähkövirta
L	sakkofunktio
P, P_h	PV-järjestelmän teho, hetken h teho
P_0	PV-järjestelmän nimellisteho
R^2	selitysaste (*coefficient of determination)*
T_a	ilman lämpötila
T_c	kennon lämpötila
T_m	paneelin lämpötila
U	jännite
ΔT	paneelin takaosan ja kennon välinen lämpötilaero
Θ	päätöspuun parametrit
$\boldsymbol{\theta}$	neuroverkkomallin parametrit
η	oppimisnopeus
γ	paneelin lämpötilakerroin
WS	tuulen nopeus
∇	gradientti

θ_T	paneelin kallistuskulma
θ_Z	auringon zeniittikulma
a, b, c, d	fysikaalisen PVUSA-mallin kertoimet
h	mittaushetki
$k_1, k_2, \ldots, k_6$	fysikaalisen 6k-mallin kertoimet
k	neuroverkon syvyys eli kerrosten lukumäärä
p_1, p_2	Perezin diffuusiomalliin liittyvät vakiotermit
s_1, s_2	Sandian kehittämän paneelin lämpötilamallin kertoimet
t, t_h	aika, mittaushetken h pituus
AOI	auringon säteilyn ja PV-paneelin välinen kulma (*angle of incidence*)
CNN	konvoluutioneuroverkko (*convolution neural network*)
DHI	diffusoitunut horisontaalisäteily (*diffused horizontal irradiance*)
DNI	suora normaalisäteily (*direct normal irradiance*)
GHI	globaali horisontaalisäteily (*global horizontal irradiance*)
HGBR	histogrammeihin perustuva gradienttiboostausmalli (*histogram gradient boosting regressor*)
IEC	kansainvälinen standardointiorgnanisaatio (*International Electrotechnical Commission*)
IL	Ilmatieteen laitos
LSTM	pitkän aikavälin lähimuistimalli (*long short-term memory*)
MAE	keski-itseisvirhe (*mean absolute error*)
MLP	monikerroksinen perseptroni (*multilayer perceptron*)
NREL	Yhdysvalloissa toimiva uusiutuvan energian tutkimuskeskus (*National Renewable Energy Laboratory*)
NWP	numeerinen säämallinnus (*numerical weather prediction*)
POA	paneelin pinnan mukainen säteily (*plane-of-array irradiance*)
PV	aurinkoenergia (*photovoltaics*)
ReLu	yksi yleinen aktivaatiofunktio (*rectified linear unit*)
RMSE	jäännösvirhehajonta (*root mean squared error*)
STC	standarditestiolosuhde (*standard test condition*)
SVM	tukivektorikone (*support vector machine*)
Turku AMK	Turun ammattikorkeakoulu

Wp	wattipiikki (*watt peak*)

1. JOHDANTO

Joissain maissa aurinkoenergia on jo valtavirtaa, kun taas jossain maissa, kuten Suomessa, aurinkoenergia on vasta nousemassa aurinkovoimateknologian kehittyessä ja halventuessa. Tällä hetkellä (maaliskuussa 2025) Suomen aurinkovoiman kokonaiskapasiteetti on noin 1260 megawattia (MW) [1], josta 123 MW on peräisin suuremmista aurinkovoimaloista, joiden kapasiteetti on yli 1 MW [2]. Tällä hetkellä siis vain noin 10 % Suomen aurinkovoimakapasiteetista tulee suuremmista voimaloista. Suomeen on kuitenkin suunnitteilla tai rakenteilla 285 uutta aurinkovoimalaa, joiden kapasiteetti on yli 1 MW. Näiden yhteenlaskettu kapasiteetti olisi yhteensä 23379 MW [3]. Tämä kymmenkertaistaisi Suomen tämänhetkisen suurten aurinkovoimaloiden (kapasiteetiltaan yli 1 MW) lukumäärän [2] ja lähes 20-kertaistaisi aurinkovoiman kokonaiskapasiteetin Suomessa. Myös pienempien aurinkovoimaloiden kapasiteetti tulee todennäköisesti lisääntymään. Jotta aurinkoenergiajärjestelmät toimisivat odotetusti, on tärkeää monitoroida niiden toimintaa. Monitorointi tarkoittaa esimerkiksi tuotetun tehon vertaamista järjestelmän teoreettiseen tehoon, eli siihen, mitä järjestelmän pitäisi kunakin hetkenä tuottaa.

Järjestelmän teoreettista tehoa voidaan estimoida tehomalleilla. Yleisesti käytettyjä tehomalleja ovat erilaiset fysikaaliset mallit, jotka ovat yksinkertaisia regressiomalleja [4]–[6]. Monimutkaisempiin tehomalleihin kuuluu erilaiset tilastolliset ja koneoppimispohjaiset mallit sekä erilaiset mallien yhdistelmät [7]. Koska tehon tuotanto on hyvin epälineaarinen ja monimutkainen prosessi, on monimutkaisemmilla malleilla parempi kyky kuvata tätä prosessia ja tuottaa tarkempia ennusteita [7]. Koneoppimismenetelmien kehittymisen myötä niitä on alettu hyödyntämään myös aurinkoenergian tehon tuotannon ennustamisessa [7], [8].

Monitoroinnilla voidaan tarkastella esimerkiksi aurinkopaneelien ikääntymistä ja toimintahäiriöitä. Tämä vaatii tarkkaa tehon mallintamista. Yleisesti käytetyt tehomallit antavat kuitenkin virheellisiä tuloksia pohjoisissa olosuhteissa etenkin talvisin heikon säteilyn ja alhaisen auringon korkeuskulman takia [6]. Myös lumi ja vaihtelevat sääolosuhtet vaikeuttavat tehon mallintamista pohjoisissa olosuhteissa [9].

Pohjoisiin olosuhteisiin sopivia tehomalleja on tutkittu aiemmin. Osassa tutkimuksista tarkastellaan tehon ennustamista tuleville ajanjaksoille [4], [7]. Näissä tutkimuksissa koneoppimismallit antavat fysikaalisia malleja parempia ennusteita. Osassa tutkimuksista taas

tarkastellaan tehon hetkittäistä mallintamista [5], [6], [10]. Näissä tutkimuksissa pääsääntöisesti keskitytään fysikaalisten mallien tarkasteluun koneoppimismallien sijasta.

Tässä työssä vertaillaan fysikaalisia malleja ja koneoppimismalleja hetkittäisen tehon mallintamiseen. Tämä työ pyrkii vastaamaan kahteen tutkimuskysymykseen:

1. Kuinka paljon yleisesti käytetyt fysikaaliset tehomallit aiheuttavat virhettä arvioituun hetkittäiseen tehoon sekä tuotettuun kokonaisenergiaan kuukausi- ja vuositasolla?
2. Voidaanko koneoppimispohjaisilla aurinkosähköjärjestelmien tehomalleilla arvioida tehon tuotantoa pohjoisissa olosuhteissa tarkemmin kuin fysikaalisilla malleilla?

Joitain fysikaalisia tehomalleja voidaan kouluttaa sopimaan tiettyyn aurinkovoimajärjestelmään paremmin, mikäli kyseisestä aurinkovoimajärjestelmästä on olemassa historiallista aineistoa. Koska uuden aurinkovoimalan valmistuessa historiallista aineistoa ei vielä ole, on tärkeää, että monitoroinnissa käytetyt mallit toimivat myös ilman kouluttamista. Tämän takia työssä käytettyjä malleja testataan aluksi suoraan uuteen järjestelmään, jonka jälkeen mallit pyritään sovittamaan järjestelmiin käyttämällä historiallista aineistoa.

2. AURINKOENERGIA

Aurinkoenergia on energiaa, jota saadaan auringon säteilystä. Se on uusiutuva energiamuoto. Aurinkoenergiasta käytetään englanniksi nimitystä photovoltaics, joka usein lyhtennetään muotoon PV. Tässäkin työssä aurinkoenergiateknologiasta puhuttaessa käytetään lyhennettä PV.

2.1 Aurinkosähköjärjestelmät

Aurinkoenergiajärjestelmät, eli PV-järjestelmät, koostuvat erilaisista komponenteista. PV-järjestelmän keskiössä on aurinkokenno, joka muuntaa auringon säteilyenergiaa sähköksi. Aurinkopaneelit koostuvat useasta aurinkokennosta, jotka on yhdistetty toisiinsa. Myös aurinkopaneeleita voidaan kytkeä toisiinsa, jotta järjestelmästä saadaan käyttötarkoitukseen sopivaa jännitettä ja virtaa. Tämä mahdollistaa aurinkoenergian hyvän skaalautuvuuden. Isojen PV-järjestelmien avulla voidaan tuottaa sähköä suoraan jakeluverkkoon. Yksityishenkilöt voivat puolestaan tuottaa sähköä omaan käyttöönsä esimerkiksi katolle asennetun aurinkopaneelin avulla. [11, luku 1]

PV-järjestelmä tuottaa jännitettä U ja virtaa I, josta voidaan laskea PV-järjestelmän tuottama hetkittäinen teho

$$P = U \cdot I. \tag{2.1}$$

Tehon yksikkö on watti W, joka puolestaan voidaan kirjoittaa myös muodossa J/s. Teho ilmaisee siis tuotetun energian määrää aikayksikössä. Aurinkopaneeleille on määritelty nimellisteho, joka on paneelin laskettu teoreettinen teho standarditestiolosuhteissa (*standard test conditions*, STC). Kansainvälisen standandardointiorganisaatio IEC (*International Electrotechnical Commission*, IEC) esittää nimellistehon yksiköksi kilowattia (kW) [12], mutta tässä työssä nimellistehon yksikkönä käytetään wattipiikkiä (*watt peak*, Wp). STC-oloissa säteily on 1000 W/m², paneelin lämpötila on 25 °C ja säteilyn ilmamassakerroin on 1,5 [11, luku 1]. Nimellistehon käyttäminen helpottaa paneelien vertailua. Järjestelmien tehon tuotannon skaalaaminen nimellisteholla helpottaa myös eri järjestelmien tuotannon vertailua.

Kun halutaan tarkastella PV-järjestelmän pidempiaikaista tuotantoa, tarkastellaan usein

järjestelmän energian tuotantoa. Energia saadaa kertomalla teho ajalla

$$E = P \cdot t. \tag{2.2}$$

Sähköenergian yksikkönä käytetään usein wattituntia (Wh), joka vastaa yhden tunnin tuotantoa 1 W teholla. PV-järjestelmien energian tuotanto lasketaan kaavalla

$$E = \sum_h P_h \cdot t_h, \tag{2.3}$$

missä t_h edustaa hetken h mittausjakson kestoa ja P_h edustaa hetkeä h vastaavaa tehoa [12]. Energian tuotantoa voidaan tarkastella esimerkiksi kuukausi- ja vuositasolla. Kun energian tuotanto skaalataan PV-järjestelmän nimellisteholla, energian tuotannon yksiköksi saadaan Wh/Wp [12]. Tämä skaalaus helpottaa erilaisten PV-järjestelmien vertailua.

PV-järjestelmän tuottama teho riippuu sääolosuhteista sekä järjestelmän kokoonpanosta. Merkittäviä energian tuotantoon vaikuttavia tekijöitä ovat säteilyn määrä sekä paneelin lämpötila. Paneelit tuottavat tasavirtaa, joka muunnetaan käyttöön sopivaksi vaihtovirraksi invertterin avulla. Aurinkopaneelien ja -kennojen tarkempi toimintaperiaate ohitetaan tässä työssä, mutta kattava kuvaus aurinkokennojen fysikaalisesta toimintaperiaatteesta esitetään esimerkiksi kirjan [13] luvussa 3. Aurinkopaneeleita rakennetaan eri materiaaleista, ja eri paneelityypeillä on erilaiset hyötysuhteet. Käytetyin materiaali aurinkopaneeleihin on pii. [11, luku 1]

Jotta PV-järjestelmä toimii suunnitellusti, on tärkeää monitoroida järjestelmän tehon ja energian tuotantoa. Erityisesti kaupallisilla aurinkovoimaloilla pienikin vika PV-järjestelmässä voi johtaa rahallisiin tappioihin [14]. Tämä johtuu siitä, että kaupallisilla aurinkovoimaloilla on usein isot PV-järjestelmät, joissa esimerkiksi useita paneeleita on kytketty toisiinsa. Jos jossain paneelissa ilmenee häiriö, se voi pahimmillaan haitata myös muiden paneelien tuotantoa, vaikka muut paneelit muuten toimisivatkin normaalisti. Pienemmillä PV-järjestelmillä voi toisaalta olla tärkeintä, että PV-järjestelmä toteuttaa varmasti sille määritellyn tehtävän, kuten jääkaapin pitämisen kylmänä ilman katkoja [14]. Järjestelmän monitorointi on tärkeää toimintahäiriöiden havaitsemiseksi molemmissa tilanteissa, vaikka PV-järjestelmän tarkoitus vaihteleekin. Yksinkertainen tapa huomata poikkeamat PV-järjestelmän toiminnassa on verrata todellista järjestelmän tuottamaa tehoa mallinnettuun tehoon [15].

Auringon säteilystä voidaan tarkastella erilaisia komponentteja. Näitä ovat globaali horisontaalisäteily (*global horizontal irradiance*, GHI), diffusoitunut horisontaalisäteily (*diffused horizontal irradiance*, DHI), suora normaalisäteily (*direct normal irradiance*, DNI), paneelin pinnan mukainen säteily (*plane-of-array irradiance*, POA). GHIn avulla tarkastellaan vaakasuoralle alueelle tulevaa kokonaissäteilyä. DHIn avulla tarkastellaan vaakasuo-

ralle alueelle tulevaa ilmakehästä siroavaa säteilyä. DNIn avulla tarkastellaan aurinkoon nähden kohtisuorassa olevalle alueelle tulevaa suoraa säteilyä. POAn avulla tarkastellaan paneelin pinnan suuntaiselle alueelle tulevaa kokonaissäteilyä. Kaikkien säteilykomponenttien yksikkö on W/m² [12].

Jokainen säteilykomponentti voidaan mitata suoraan mittalaitteella tai ne voidaan laskea matemaattisesti muiden säteilykomponenttien avulla. Erityisen tärkeä säteilykomponentti on POA-säteily, koska se ilmoittaa suoraan paneelille tulevan säteilyn. POA-säteilyä voidaan mitata pyranometrilla. Pyranometri on kuitenkin monimutkainen ja kallis mittalaite, joten sitä ei ole aina saatavilla PV-järjestelmien yhteydessä. Etenkin pienemmillä PV-järjestelmillä ei välttämättä ole tarvetta tarkalle POA-säteilyn mittaamiselle. POA-säteily voidaan kuitenkin laskea myös muiden säteilykomponenttien avulla [16]

$$G_{POA} = G_b + G_g + G_d, \text{ jossa} \tag{2.4}$$

$$G_b = DNI \cdot \cos(AOI) \tag{2.5}$$

$$G_g = GHI \cdot albedo \cdot \frac{1 - \cos(\theta_T)}{2} \tag{2.6}$$

$$G_d = DHI \cdot \frac{1 + \cos(\theta_T)}{2}. \tag{2.7}$$

Kaava 2.5 viittaa säteilyn suoraan komponenttiin, jossa AOI (*angle of incidence*) tarkoittaa kulmaa auringon säteilyn ja PV-paneelien välillä. Kaava 2.6 viittaa säteilyn maasta heijastuneeseen komponenttiin. Siinä albedo tarkoittaa maan heijastavuutta. Pienempi albedon arvo tarkoittaa pienempää maan heijastavuutta ja suurempi arvo tarkoittaa maan suurempaa heijastavuutta. Kaava 2.7 viittaa säteilyn sironneeseen eli diffusoituneeseen komponenttiin. Kaavassa 2.7 esiintyvä kulma θ_T tarkoittaa paneelin kallistuskulmaa. Kaavaa 2.7 kutsutaan isotrooppiseksi diffuusiomalliksi ja sitä käytetään usein pohjana monimutkaisemmille diffuusiomalleille, kuten Perezin diffuusiomallille [17]

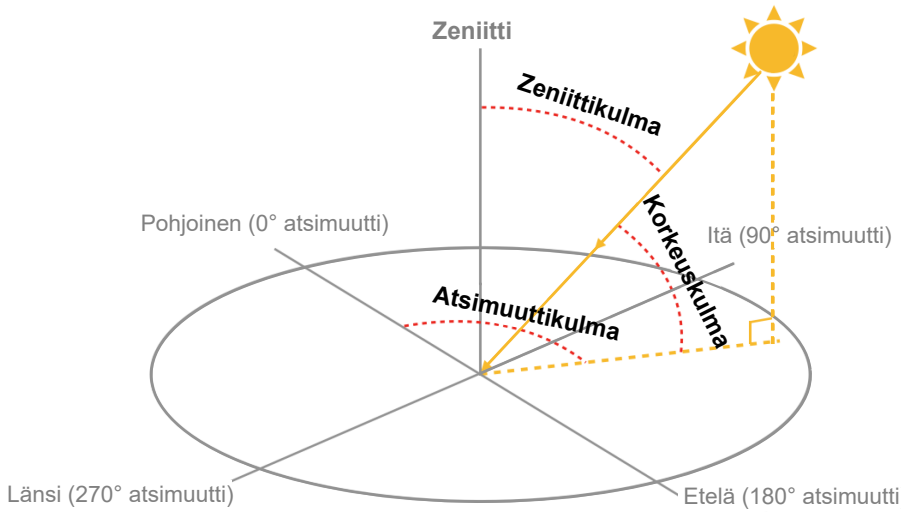
$$G_d = DHI \cdot \left((1 - F_1) \left(\frac{1 + \cos(\theta_T)}{2} \right) + F_1 \left(\frac{p_1}{p_2} \right) + F_2 \sin(\theta_T) \right), \tag{2.8}$$

missä F_1 ja F_2 ovat empiirisesti muodostettuja malleja, jotka kuvaavat säteilystä johtuvaa kirkkautta. Vakiotermit p_1 ja p_2 on puolestaan määritelty

$$p_1 = \max(0, \cos(AOI)) \tag{2.9}$$

$$p_2 = \max(\cos(85°), \cos(\theta_Z)), \tag{2.10}$$

missä θ_Z tarkoittaa auringon zenittiikulmaa (*zenith angle*), joka ilmaisee auringon kulman pystysuoraan tasoon nähden. Toinen tärkeä myöhemmin vastaan tuleva kulma on atsimuuttikulma (*azimuth angle*), joka ilmaisee auringon tai paneelin kulmaa pohjoisesta myötäpäivään nähden. Kuvassa 2.1 on esitetty auringon zeniitti- ja atsimuuttikulmat sekä aurignon korkeuskulma.



Kuva 2.1. *Auringon zeniitti-, atsimuutti- ja korkeuskulma. Myös paneelin atsimuuttikulma määritetään samalla tavalla pohjoisesta myötäpäivään. Kuva muokattu lähteestä [18] (CC-BY 4.0).*

Aina paneelien lämpötilaa ei pystytä mittaamaan, jolloin paneelien lämpötila pitää laskea sääolosuhteista. Paneelin lämpötilaa T_m voidaan estimoida erilaisilla malleilla. Tässä työssä käytetään Sandian kehittämää paneelin lämpötilamallia [19]. Siinä lämpötila T_m lasketaan POA-säteilyn G_{POA} ja tuulen nopeuden WS avulla

$$T_m = G_{POA} \cdot \left(e^{s_1 + s_2 \cdot WS}\right) + T_a. \tag{2.11}$$

Tuulen nopeus on mitattu 10 metrin korkeudella yksikössä m/s, T_a on ilman lämpötila celsius-asteissa ja s_1 ja s_2 ovat kokeellisia kertoimia, joille löytyy valmiiksi lasketut arvot eri paneelimateriaaleille ja asennuksille.

Jotkin tehomallit tarvitsevat syötteeksi paneelin lämpötilan sijaan kennon lämpötilan T_c. Se voidaan laskea esimerkiksi paneelin lämpötilan avulla. Tässä työssä käytetään Sandian kehittämää mallia [19]

$$T_c = T_m + \frac{G_{POA}}{G_{POA,\mathrm{STC}}} \cdot \Delta T, \tag{2.12}$$

missä $G_{POA,\mathrm{STC}}$ on STC-olosuhteiden mukainen POA-säteily, jonka arvo on 1000 W/m² ja ΔT on paneelin takaosan ja kennon välinen lämpötilaero. Lämpötilaero vaihtelee 0 ja 3 asteen välillä ja se riippuu paneelin materiaalista ja asennuksesta.

2.2 Tehon mallintaminen

Koska PV-järjestelmän teho riippuu sääolosuhteista, ottavat tehomallit usein parametriksi erilaisia säähavaintoja. Jotkut PV-järjestelmät saattavat sisältää oman säähavaintoaseman tarkkojen havaintojen tekemiseksi, mutta suurella osalla PV-järjestelmistä ei ole

omaa säähavaintoasemaa. Etenkin pienemmillä PV-asemilla ei todennäköisesti ole kattavaa sääasemaa. Tällöin tehomallin tarvitsemat säähavainnot voidaan laskea toisista säähavainnoista matemaattisilla kaavoilla, kuten 2.4, 2.11 ja 2.12. Sääolosuhteita voidaan mitata myös esimerkiksi satelliittien avulla, mikä mahdollistaa sääolosuhteiden mittaamisen myös syrjäisille sijainneille, joissa ei ole lähistöllä sääasemaa. Sääolosuhteita voidaan myös estimoida numeerisilla säämallinnuksilla (*numerical weather prediction*, NWP) [20].

Tehomalleja voidaan jaotella mallin tyypin mukaan esimerkiksi fysikaalisiin, tilastollisiin, koneoppimispohjaisiin ja hybridimalleihin. Fysikaaliset mallit laskevat tehon erilaisten PV-järjestelmän parametrien ja säähavaintojen avulla. Tilastolliset menetelmät olettavat havaintojen välillä olevan ajallista riippuvuutta ja käyttävät ennusteen luomiseen historiallista dataa. Käytettyjä tilastollisia malleja ovat erilaiset aikasarjamallit. Koneoppimispohjaiset mallit pyrkivät oppimaan tehon tuotannon tilastollisten mallien tapaan historiallisen datan avulla, mutta ne eivät välttämättä oleta havaintojen välillä olevan ajallista riippuvuutta. Esimerkkeinä koneoppimismalleista voidaan pitää tukivektorikonetta, satunnaismetsää ja neuroverkkoa. Hybridimallit yhdistelevät erilaisia malleja keskenään. Tällainen jaottelu ei ole kuitenkaan yksiselitteinen, ja malleja onkin jaoteltu eri tavoilla. [7]

Koska PV-järjestelmän tehon tuotanto riippuu monesta tekijästä, on vaikea mallintaa tehoa luotettavasti [15]. Koneoppimismenetelmien kehittymisen myötä näitä malleja on alettu hyödyntämään myös PV-järjestelmien tehon tuotannon mallintamisessa [8], ja ne ovatkin nykyään suosituin menetelmä siihen [7]. Esimerkiksi erilaiset neuroverkkomallit ovat antaneet hyviä tuloksia tehon ennustamisessa johtuen osin siitä, että neuroverkkomallit pystyvät mallintamaan tehon tuotantoon liittyviä epälineaarisia riippuvuuksia hyvin [7].

Tehon mallintamiseen liittyy läheisesti myös tehon ennustaminen, jossa reaaliaikaisen tehon sijasta ennustetaan tehon tulevia arvoja. Vaikka tässä työssä keskitytään tehon mallintamiseen ennustamisen sijaan, iso osa aiemmista tutkimuksista keskittyy tehon ennustamiseen eri ennustehorisonteille (*forecast horizon*). Kun aihetta rajataan käsittämään vain pohjoiset olosuhteet, aiempia tutkimuksia tehon mallintamisesta löytyy varsin vähän. Tämän takia tässä työssä tarkastellaan myös tutkimuksia, joissa ennustetaan tehoa eri ennustehorisonteille.

2.3 Pohjoisten olosuhteiden vaikutus tehon mallintamiseen

Pohjoiset olosuhteet tarjoavat aurinkovoimalle hyötyjä ja haittoja. Aurinkokennojen hyötysuhde kasvaa kylmemmissä olosuhteissa, mikä parantaa sähkötehon tuotantoa [7]. Myös lumi voi heijastaa osan auringon säteilystä aurinkopaneeleihin, jolloin säteilyä päätyy paneeliin enemmän ja tehon tuotanto kasvaa [7]. Aurinkopaneelien on myös näytetty ikääntyvän pohjoisissa olosuhteissa hitaammin kuin eteläisemmissä olosuhteissa [21]. Lisäksi kesän pitkät päivät tarjoavat säteilyä ison osan vuorokaudesta. Toisaalta tehon tuotantoa

laskee matalampi auringon korkeuskulma ja talvisin lumi sekä lyhyet päivät [7]. Matalasta auringon tulokulmasta johtuen myös varjostuminen voi koitua ongelmaksi, mikäli PV-järjestelmän ympärillä on esimerkiksi korkeita rakennuksia tai puita. Erityisesti lumi ja jää ovat hyvin vaikeasti mallinnettavia vaikka näitä ollaankin tutkittu. On kuitenkin arvioitu, että sähköntuotannon häviö lumen takia on yleensä vähemmän kuin 10 % vuodessa, mutta talvikuukausina lumeen liittyvä sähköhäviö voi kuitenkin olla yli 25 % kuukausitasolla [22]. Lumesta on siis hyötyä silloin, kun se heijastaa maahan osuvaa säteilyä paneeliin, mutta haittaa silloin, kun paneeli on peittynyt lumikerrokseen, joka estää säteilyn osumista paneeliin.

Herman Böök tarkasteli ja kehitti PV-järjestelmiin liittyviä malleja ja työkaluja väitöskirjassaan [23]. Erityisesti väitöskirjaan liittyvät julkaisut [5] ja [24] liittyvät läheisesti tämän diplomityön aiheeseen. Julkaisussa [5] Böök tarkasteli fysikaalisen 6k-mallin toimintaa Suomen oloissa. Aineistona hän käytti Helsingissä ja Kuopiossa sijaitsevia Ilmatieteen laitoksen PV-järjestelmiä, joita myös tässä työssä käytetään. Kumpaankin kohteeseen sovitettiin myös oma malli käyttäen osaa aineistosta. Julkaisussa [24] Böök täsmäkoulutti julkaisussa [5] kehittämäänsä mallia 23 eri PV-järjestelmään. Koska sääaineisto oli NWP-aineistoa, oli se aineistosta suurin epävarmuuden aiheuttaja. NWP-aineiston käyttäminen mahdollistaa toisaalta mallin laajemman käytettävyyden erityisesti järjestelmille, joiden läheisyydessä ei ole kattavaa sääasemaa.

Karttunen ym. vertailivat tutkimuksessa [6] eri fysikaalisia malleja Turussa sijaitsevan PV-järjestelmän ikääntymisen tarkasteluun. PV-järjestelmä sisälsi kaksipuoleisia pystysuoraan asennettuja paneeleita, jotka oltiin asennettu itä-länsi-suuntaisesti. Tutkimuksessa huomattiin, että vaikka osassa malleista oli lämpötilakorjaus, jonka pitäisi vähentää säästä johtuvaa kausittaista vaihtelua kesän ja talven välillä, tuottivat kaikki mallit silti poikkeavia arvoja talvisin. Pahimmillaan lämpötilakorjaus lisäsi mallin kausittaista vaihtelua sen sijaan, että vaihtelu olisi vähentynyt.

Brester ym. vertailivat tutkimuksessa [4] koneoppimismalleja ja fysikaalista 6k-mallia tehon ennustamiseen. Työssä tarkasteltiin tehon ennustamista seuraavalle päivälle. Tutkimuksessa käytettiin puolen vuoden mittaista aineistoa kolmesta eri järjestelmästä Kuopiossa. Yksi tarkastelluista järjestelmistä oli Ilmatieteen laitoksen PV-järjestelmä, jota myös tässä työssä käytetään. Tutkimuksen mukaan koneoppimismallit paransivat ennustustarkkuutta fysikaalisiin malleihin verrattuna.

Dimd ym. tarkastelivat kirjallisuuskatsauksessa [7] kirjallisuudessa esiintyviä koneoppimismenetelmiä tehon tuotannon ennustamiseen pohjoisissa olosuhteissa. Kirjallisuuskatsaus keskittyi erityisesti Norjan olosuhteisiin. Koska sää voi muuttua pohjoisissa olosuhteissa nopeastikin esimerkiksi pilvien muodostuessa, suosittelevat kirjoittajat erityisesti lyhyen ennustehorisontin käyttämistä pohjoisissa oloissa. He eivät kuitenkaan käyneet läpi tehon mallintamiseen liittyviä tutkimuksia. He suosittelevat lumesta johtuvan paneelin

likaantumisen ottamista huomioon ennustemallien tekemisessä.

Awad ym. tutkivat tutkimuksessa [25] PV-järjestelmän päivittäisen energian tuotannon ennustamista lumisissa olosuhteissa. Tutkimus keskittyi Kanadassa sijaitseviin PV-järjestelmiin. Tutkimuksen tavoitteena oli ottaa paneelien likaantuminen ja erityisesti lumipeite huomioon ennustuksessa. Tutkimuksen aineisto sisälsi 85 aurinkosähkön pientuottajaa erilaisilla sijainneilla ja paneelien asennuksilla. Energian ennustamiseen käytettiin neuroverkkomallia. Tutkimuksessa huomattiin, että paneelit kannattaa asentaa 50 tai 60 asteen kallistuskulmaan, koska korkeampi kallistuskulma antaa pohjoisissa olosuhteissa energiaa tasaisemmin ympäri vuoden ja paneelit ovat tällöin myös vähemmän alttiita lumipeitteelle. Tutkimuksessa suositeltiin erillisten mallien tekemistä kesälle ja talvelle.

Jouttijärvi ym. tutkivat tutkimuksessa [10] kaksipuoleisten aurinkopaneelien energian tuotannon ennustamista eri ennustehorisonteille. Tutkimuksessa tehon mallintamiseen käytettiin fysikaalista 6k-mallia ja yksinkertaisempaa säteilyyn pohjautuvaa tehomallia. Tuloksista ilmenee, että 6k-malli ennustaa energian tuotantoa tarkemmin kuin yksinkertaisempi säteilyyn pohjautuva tehomalli. Myös laadukas data ja datan aggregointi vähentää ennusteen virhettä. Tutkimuksessa suositellaan säteilyn mittaamista PV-järjestelmän lähistöltä etenkin korkearesoluutioisen aineiston yhteydessä.

Pohjoisille olosuhteille tyypillinen kausittainen vaihtelu vaikeuttaa tehon mallintamista ja ennustamista. Fysikaaliset mallit näyttäisivät joidenkin tutkimusten perusteella toimivan suomen oloissa hyvin [5], [24], mutta niiden on osoitettu antavan virheellisiä tuloksia erityisesti talvisin [6]. Koneoppimismallit näyttäisivät soveltuvan hyvin etenkin tehon ennustamiseen tuleville ajanhetkille [7], [25] ja antavan fysikaalisia malleja tarkempia tehoennusteita [4], mikä motivoi niiden käyttöä myös tehon mallintamiseen.

2.4 Fysikaaliset mallit

Fysikaalisilla malleilla tarkoitetaan kirjallisuudessa välillä eri asioita. Sillä voidaan tarkoittaa esimerkiksi malleja, jotka estimoivat tehon PV-kennon ekvivalenttivirtapiiriä (*equivalent electric circuit*) kuvaavilla yhtälöillä [26] tai malleja, jotka estimoivat tehon empiirisillä regressiomalleilla [4]. Tässä työssä fysikaalisilla malleilla tarkoitetaan empiirisiä regressiomalleja, jotka ennustavat tehon esimerkiksi säteilyn ja paneelin lämpötilan avulla. Tässä työssä tarkastellaan fysikaalisia malleja, joita on käytetty tutkimuksessa [6].

2.4.1 6k

6k on regressiomalli, joka estimoi tuotetun tehon POA-säteilystä, paneelin lämpötilasta ja paneelin nimellistehosta

$$P(G_{POA}, T_m, P_0) = G'_{POA}(P_0 + k_1 \ln(G'_{POA}) + k_2 \ln(G'_{POA})^2 + k_3 T' + k_4 T' \ln(G'_{POA}) + k_5 T' \ln(G'_{POA})^2 + k_6 T'^2), \tag{2.13}$$

missä $G'_{POA} = G_{POA}/G_{POA,\mathrm{STC}}$ ja $T' = T_m - T_{\mathrm{STC}}$ [27]. Termi T_{STC} on STC-olosuhteiden mukainen lämpötila, jonka arvo on 25 °C. Kaavassa 2.13 P on mallinnettu teho, P_0 on paneelin nimellisteho ja $k_1, k_2, \ldots, k_6$ ovat mallin kertoimia, jotka ratkaistaan sovittamalla osa aineistosta malliin. 6k-malli voidaan skaalata järjestelmän nimellisteholla, jolloin eri järjestelmiin sovitettujen mallien kertoimia on helpompi vertailla keskenään. Tällöin

$$P'(G_{POA}, T_m) = G'_{POA}(1 + k'_1 \ln(G'_{POA}) + k'_2 \ln(G'_{POA})^2 + k'_3 T' + k'_4 T' \ln(G'_{POA}) + k'_5 T' \ln(G'_{POA})^2 + k'_6 T'^2), \tag{2.14}$$

missä $P' = P/P_0$ ja $k'_1 = k_1/P_0, k'_2 = k_2/P_0, \ldots, k'_6 = k_6/P_0$ [27]. 6k-mallia on aikaisemmin käytetty Suomen olosuhteissa [4]–[6], [10] ja kertoimien sovituksen jälkeen se on tutkimuksissa [10] ja [6] käytetyistä malleista tarkin.

2.4.2 PVWatts

PVWatts on Yhdysvalloissa toimivan uusiutuvan energian tutkimuskeskuksen NRELin (*National Renewable Energy Laboratory*) kehittämä yksinkertainen malli, joka estimoi tuotetun tehon POA-säteilystä ja kennon lämpötilasta

$$P(G_{POA}, T_c, \gamma, P_0) = \frac{C \cdot P_0 \cdot G_{POA}}{G_{POA,\mathrm{STC}}}, \tag{2.15}$$

missä C on lämpötilakorjaustermi $C = 1 + \gamma(T_c - T_{\mathrm{STC}})$. Lämpötilakorjaustermissä C termi γ on paneelin lämpötilakerroin. Koska mallissa ei ole vakiokertoimia, mallia ei tarvitse kouluttaa ja se sopiikin hyvin nopeaan tehon ennustamiseen ilman aiempaa aineistoa. Sen virhe voi kuitenkin olla vuositasolla jopa ±20 % ja kuukausitasolla ±40 %. [28]

2.4.3 PVUSA

PVUSA on yksinkertainen regressiomalli, joka estimoi tuotetun tehon POA-säteilystä, ilman lämpötilasta ja tuulen nopeudesta

$$P(G_{POA}, T_a, WS) = G_{POA} \cdot (a + b \cdot G_{POA} + c \cdot T_a + d \cdot WS), \tag{2.16}$$

missä a, b, c, d ovat mallin kertoimia, jotka ratkaistaan sovittamalla osa aineistosta malliin [29]. Mallin heikkouksina on mainittu huono toimivuus alhaisilla säteilyillä, joten on suositeltu, että mallia ei käytetä havaintoihin, joissa säteily on alle 500 W/m². Paneelille osuva säteily riippuu siitä, onko paneeli yksi- vai kaksipuolinen ja miten paneeli on asennettu. Tutkimuksessa [6] tutkittiin kaksipuoleisia pystysuoraan asennettuja paneeleita Suomessa ja suurin osa paneeleille osuvasta säteilystä oli alle 200 W/m². PVUSA-malli tuottikin tutkimuksessa epätarkkoja tehoennusteita.

3. KONEOPPIMISMALLIT

Koneoppimismallit ovat malleja, jotka oppivat aineistosta. Yleisesti tämä tarkoittaa sitä, että mallin suorituskyky paranee määrätyssä tehtävässä mallin saadessa lisää kokemusta aineiston avulla. Esimerkkeinä koneoppimismalleille sopivista tehtävistä ovat regressio- ja luokittelutehtävät. Regressiotehtävissä pyritään ennustamaan numeerisen muuttujan arvoa ja luokittelutehtävissä pyritään ennustamaan kategorisen muuttujan kategoriaa. Regressio- ja luokittelutehtävien ratkaisemiseen voidaan käyttää ohjattua oppimista (*supervised learning*), jossa malli koulutetaan aineistolla, jossa havaintojen lisäksi saatavilla on myös ennustettavan muuttujan numeerinen arvo (regressiotehtävät) tai kategoria (luokittelutehtävät) [30, s. 97–103]. Koska tässä työssä pyritään estimoimaan PV-järjestelmän tehoa, joka on numeerinen muuttuja, on kyseessä regressiotehtävä.

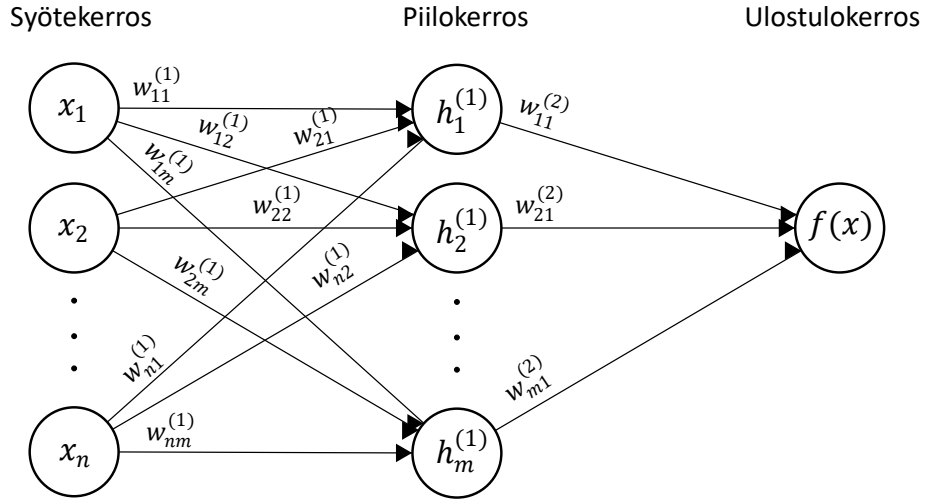
Tehon estimointiin voidaan käyttää erilaisia koneoppimismalleja. Esimerkiksi tukivektorikoneet (*support vector machines*, SVM), koostepuumallit (*ensemble of trees*), monikerroksiset päättelimet (*multilayer percpetrons*, MLP), konvoluutioneuroverkot (*convolution neural networks*, CNN) ja pitkän aikavälin lähimuistimallit (*long short-term memory*, LSTM) ovat hyvin käytettyjä etenkin tehon ennustamisessa [7]. Tässä työssä tarkastellaan kahta erilaista koneoppimismallia. Aluksi tarkastellaan MLP-mallia, jonka jälkeen tarkastellaan koostepuumalleihin kuuluvaa gradienttiboostausmallia.

3.1 Monikerroksinen päättelin

Monikerroksinen päättelin, eli MLP, on myötävirtaneuroverkko (*feed-forward neural network*), joka määritellään kuvauksena

$$y = f(\mathbf{x}; \boldsymbol{\theta}). \tag{3.1}$$

Myötävirtaneuroverkoissa informaatio liikkuu syötteistä $\mathbf{x}$ funktion f kautta ennusteeseen y. Kaavassa 3.1 $\boldsymbol{\theta}$ tarkoittaa MLP-mallin parametreja, jotka pyritään optimoimaan. MLP-mallit koostuvat useasta funktiosta, jotka on yhdistetty toisiinsa. MLP-mallia voisi siis kuvata esimerkiksi funktioketjulla $f(\mathbf{x}) = f_3(f_2(f_1(\mathbf{x})))$, missä f_1 on neuroverkon ensimmäinen kerros, f_2 neuroverkon toinen kerros ja f_3 kolmas kerros. Funktiota f_1 kutsutaan myös syötekerrokseksi (*input layer*), funktiota f_2 kutsutaan piilokerrokseksi (*hidden layer*) ja funktiota f_3 kutsutaan ulostulokerrokseksi (*output layer*). Ensimmäinen piilokerros voi-



Kuva 3.1. *Yksinkertaisen MLP-mallin rakenne. Mallissa on n syötemuuttujaa, yksi piilokerros, joka on m neuronin levyinen, ja yksi yhden neuronin ulostulokerros.*

daan ilmaista syötteiden $\mathbf{x}$ avulla yhtälöllä

$$\mathbf{h}^{(1)} = g^{(1)}\left(\mathbf{W}^{(1)T}\mathbf{x} + \mathbf{b}^{(1)}\right), \tag{3.2}$$

jonka jälkeen muut piilokerrokset voidaan laskea edellisten piilokerrosten avulla yhtälöillä

$$\mathbf{h}^{(2)} = g^{(2)}\left(\mathbf{W}^{(2)T}\mathbf{h}^{(1)} + \mathbf{b}^{(2)}\right) \tag{3.3}$$

$$\vdots$$

$$\mathbf{h}^{(k)} = g^{(k)}\left(\mathbf{W}^{(k)T}\mathbf{h}^{(k-1)} + \mathbf{b}^{(k)}\right). \tag{3.4}$$

Kaavassa 3.4 funktio $g^{(k)}$ tarkoittaa kerroksen k aktivaatiofunktiota, matriisi $\mathbf{W}^{(k)T}$ tarkoittaa kerroksen k painoja (*weights*), vektori $\mathbf{h}^{(k-1)}$ tarkoittaa edellisen kerroksen ulostuloja ja vektori $\mathbf{b}^{(k)}$ tarkoittaa kerroksen k vakiotermejä (*bias-term*). Funktioketjun pituus, eli kerrosten lukumäärä k, määrittää neuroverkkomallin syvyyden. Piilokerrosten koko määrittää neuroverkon leveyden. Kuvassa 3.1 on esitetty yksinkertaisen MLP-mallin rakenne. Kuvassa näkyy syötekerros, jossa on n syötettä, yksi piilokerros, jossa on m neuronia, ja ulostulokerros. Kuvan mukaisen MLP-mallin aktivaatio $h_1^{(1)}$ saadaan laskemalla

$$h_1^{(1)} = g^{(1)}\left(w_{11}^{(1)} \cdot x_1 + w_{21}^{(1)} \cdot x_2 + \cdots + w_{n1}^{(1)} \cdot x_n + b_1^{(1)}\right), \tag{3.5}$$

missä vakiotermi $b_1^{(1)} = 0$, sillä vakiotermejä ei ole kuvan MLP-mallissa. Vastaavalla tavalla saadaan laskettua piilokerroksen muut aktivaatiot ja matriisimerkintää hyödyntämällä aktivaatiot voidaan kirjoittaa muotoon 3.2. Neuroverkon arkkitehtuurilla tarkoitetaan neuroverkon rakennetta, kuten neuroverkon syvyyttä, kerrosten leveyttä ja kerroksien välistä yhteyttä toisiinsa. [30, s. 164–194]

Neuroverkoissa yleisesti käytetty aktivaatiofunktio on ReLu (*rectified linear unit*), joka on muotoa

$$g(z) = \max\{0, z\}. \tag{3.6}$$

Aktivaatiofunktiot mahdollistavat mallin epälineaarisuuden. Näiden epälineaaristen aktivaatiofunktioiden ansiosta neuroverkoilla voidaan mallintaa mitä tahansa funktiota. Tätä kutsutaan yleiseksi approksimaatioteoriaksi (*universal approximation theorem*). Vaikka teoriassa mikä tahansa funktio olisi mallinnettavissa, ei neuroverkkomalli välttämättä opi tätä funktiota koulutusvaiheessa. Vaikka yleinen approksimaatioteoria kertoo, että riittävän isolla neuroverkolla pystytään mallintamaan mitä vain funktiota miten tarkasti tahansa, se ei kerro kuinka iso neuroverkon tulisi olla. [30, s. 187–195]

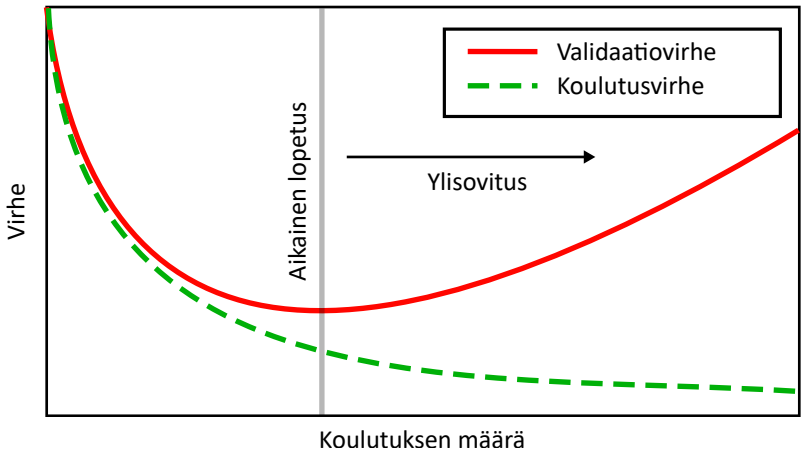
Neuroverkot koulutetaan usein iteratiivisilla gradientteihin perustuvilla algoritmeillä, jotka pyrkivät minimoimaan valittua sakkofunktiota (*cost function*) L. Gradientti saadaan laskemalla sakkofunktion L osittaisderivaatta kunkin MLP:n parametrin θ_j suhteen

$$\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) = \begin{bmatrix} \frac{\partial}{\partial \theta_1} L(\boldsymbol{\theta}) \\ \frac{\partial}{\partial \theta_2} L(\boldsymbol{\theta}) \\ \vdots \\ \frac{\partial}{\partial \theta_n} L(\boldsymbol{\theta}) \end{bmatrix}, \tag{3.7}$$

missä n on parametrien θ_j lukumäärä. Koska gradientti osoittaa suuntaan, jossa sakkofunktion L arvo kasvaa nopeiten, voidaan sakkofunktion arvo minimoida kulkemalla gradienttia vastakkaiseen suuntaan. Tätä kutsutaan gradienttimenetelmäksi (*gradient descent*). Parametrien $\boldsymbol{\theta}$ arvoa päivitetään kulkemalla gradienttia vastakkaiseen suuntaan oppimisnopeuden η verran

$$\boldsymbol{\theta}' = \boldsymbol{\theta} - \eta \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}). \tag{3.8}$$

Koska gradienttimenetelmä 3.8 käyttää parametrien optimointiin koko aineistoa, on se hidas etenkin suurilla aineistoilla. Tämän takia parametrien optimointiin käytetään nopeampia numeerisia menetelmiä [31, luku 4]. Tällainen on esimerkiksi Adam-menetelmä, joka on stokastinen momentteihin perustuva optimointialgoritmi [32]. Momenttien tarkoitus on ikään kuin pitää yllä gradienttimenetelmän liikemäärää ja vauhdittaa optimointia [31]. Stokastisuus tarkoittaa optimoinnin satunnaisuutta, joka johtuu siitä, että gradienttimenetelmästä 3.8 poiketen Adam-menetelmä optimoi parametreja $\boldsymbol{\theta}$ vain osalla aineistosta kerrallaan [32]. Näitä osia kutsutaan eriksi (*mini-batch*). Erien koko vaikuttaa mallien koulutukseen. Suuremmat eräkoot antavat tarkemman estimaatin gradientista, mutta ovat hitaampia laskea. Tyypilliset eräkoot ovat lukujen 32 ja 256 välissä [30, s. 276], mutta esimerkiksi tutkimuksessa [33] neuvotaan käyttämään eräkokoja väliltä 2 ja 32. Myös



Kuva 3.2. *Koulutus- ja validaatiovirheen muutokset koulutuksen lisääntyessä. Koulutusvirhe pienenee, mutta validaatiovirhe alkaa jossain koulutuksen vaiheessa kasvamaan, koska malli alkaa ylisovittumaan koulutusaineistoon. Aikaisessa lopetuksessa mallin koulutus lopetetaan silloin, kun validaatiovirhe on pienimmillään.*

tutkimus [34] mainitsee hyväksi eräkooksi 32. Pienet eräkoot saattavat vähentää ylisovittamisen (*overfit*) riskiä [30, s. 276]. Kun koko aineisto on käytetty kerran mallin kouluttamiseen, kutsutaan sitä epokiksi (*epoch*) [30, s. 243].

Ylisovittaminen tarkoittaa sitä, että malli sopii hyvin koulutusaineistoon, muttei uuteen aineistoon [30, s. 110]. Koska koneoppimismalleilla pyritään luomaan malli, joka toimii hyvin myös uudella aineistolla, ylisovittamista pyritään välttämään. Malli koitetaan siis kouluttaa niin, että se yleistyy (*generalize*) hyvin. Mallin yleistymistä tarkastellaan laskemalla mallin testivirhe (*test error*) aineistolla, jota ei ole käytetty mallin kouluttamiseen. Joskus testivirhettä halutaan minimoida koulutusvirheen kustannuksella. Tätä kutsutaan mallin sääntelyksi (*regularization*). Aineisto voidaan jakaa myös validaatioaineistoon (*validation set*), jota käytetään mallin yleistymisen tarkasteluun mallin koulutusvaiheessa. Validaatiovirheellä (*validation error*) estimoidaan siis mallin testivirhettä koulutuksen aikana ja testivirhettä tarkastellaan vasta, kun mallin koulutus on valmis. Kun mallia aletaan kouluttamaan, koulutusvirhe ja validaatiovirhe (ja samalla testivirhe) pienenevät. Koulutusvirhe jatkaa pienenemistä, mutta validaatiovirhe saattaa jossakin koulutuksen vaiheessa alkaa kasvamaan, kun malli alkaa sovittumaan koulutusaineistoon liian hyvin. Koulutusvirheen väheneminen ja validaatiovirheen nouseminen sekä ylisovittamisen alkaminen näkyvät kuvasta 3.2. Yksi tehokas ja yksinkertainen sääntelykeino on aikainen lopetus (*early stopping*), jossa mallin koulutus lopetetaan silloin, kun validaatiovirhe on pienimmillään. Ei ole kuitenkaan varmaa onko testivirhe tällöin pienimmillään, mutta koska testiaineistoa käytetään vain lopullisen mallin toiminnan tarkasteluun, aikainen lopetus tehdään validaatiovirheen perusteella. [30, s. 224–243]

Oppimisnopeus η on yksi tärkeimpiä mallin koulutukseen vaikuttavia tekijöitä. Sen sijaan, että oppimisnopeus olisi sama kaikille mallin parametreille, voidaan se asettaa erik-

seen jokaiselle parametrille. Lisäksi oppimisnopeuksia voidaan vaihtaa koulutuksen aikana. Tätä kutsutaan adaptiiviseksi oppimisnopeudeksi (*adaptive learning rate*). Esimerkiksi Adam-menetelmä käyttää adaptiivista oppimisnopeutta. [30, s. 302–306]

Yleensä syötemuuttujat skaalataan, jotta ne ovat suurusluokaltaan samanlaisia. Ilman skaalausta koneoppimismallit saattavat painottaa muuttujia, jolla on suurempi suurusluokka. Ulostulomuuttujaa ei yleensä tarvitse skaalata. Muuttujat voidaan skaalata esimerkiksi standardoimalla ne. Standardoinnissa muuttujista vähennetään aluksi muuttujan keskiarvo, jonka jälkeen muuttujat jaetaan niiden keskihajonnalla. Koska aineisto jaetaan koulutus- ja testiaineistoon, on kummankin aineiston muuttujat skaalattava erikseen. Tämä tapahtuu esimerkiksi standardoimalla aluksi koulutusaineiston muuttujat ja standardoimalla testiaineiston muuttujat käyttäen koulutusaineiston muuttujien keskiarvoja ja keskihajontoja. Kategorisille muuttujille tehdään usein yksi-kuuma-koodaus (*one-hot encoding*), jossa jokaista muuttujan kategoriaa kohden tehdään uusi muuttuja, joka voi saada arvoksi vain 1 tai 0. Jokaista alkuperäisen muuttujan kategoriaa vastaava uusi muuttuja saa arvoksi 1, kun alkuperäinen muuttuja saa arvokseen tämän kategorian. Muutoin uudet muuttujat saavat arvokseen 0. [31]

Kun koneoppimismallia koulutetaan, tulee valita esimerkiksi mallin arkkitehtuuri, oppimisnopeus ja eräkoko. Näitä kutsutaan mallin hyperparametreiksi. Oleellinen osa koneoppimismallien optimointia on hyperparametrien valinta, joka voidaan tehdä manuaalisesti tai automaattisesti. Manuaalisessa valinnassa hyperparametreja muokataan mallin toiminnan mukaan. Tämä vaatii hyvää ymmärrystä siitä, miten eri hyperparametrit vaikuttavat mallin toimintaan. Automaattisessa valinnassa hyperparametit valitaan käymällä kaikki vaihtoehdot läpi ja valitsemalla paras hyperparametrien yhdistelmä, joten se on laskennalisesti usein manuaalista hienosäätöä raskaampi. Eräs menetelmä automaattiseen hyperparametrien valintaan on ruutuetsintä (*grid search*). Ruutuetsinnässä kullekin hyperparametrille valitaan tietyt arvot, joita halutaan kokeilla. Koska ruutuetsinnässä kokeillaan kaikki mahdolliset hyperparametrien arvojen yhdistelmät, kokeiltavien yhdistelmien lukumäärä kasvaa eksponentiaalisesti. Ruutuetsinnästä voi tämän vuoksi tulla helposti laskennallisesti raskas. [30, s. 422–429]

MLP-mallin kouluttamisen jälkeen sitä voidaan käyttää uusien havaintojen estimointiin. MLP-mallin koulutusta voidaan kuitenkin jatkaa uudella aineistolla. Esimerkiksi yleinen PV-järjestelmän tehoa mallintava MLP-malli voidaan sovittaa tiettyyn PV-järjestelmään jatkamalla MLP-mallin kouluttamista tämän tietyn järjestelmän aineistolla. Tätä kutsutaan tässä työssä täsmäkoulutukseksi.

3.2 Gradienttiboostaus

Boostauksessa usea yksinkertainen malli yhdistetään yhdeksi malliksi. Yksinkertaisia malleja kutsutaan usein heikoiksi oppijoiksi (*weak learner*) ja yhdistettyä mallia vahvaksi op-

pijaksi (*strong learner*). Boostauksessa heikkoja oppijoita koulutetaan peräkkäin niin, että seuraava heikko oppija pyrkii oppimaan edellisen heikon oppijan virheistä [31]. Heikkona oppijana voidaan käyttää päätöspuita (*decision trees*) [35, s. 353]. Vahva oppija voidaan esittää heikkojen oppijoiden summana

$$f_M(\mathbf{x}) = \sum_{m=1}^{M} H(x; \Theta_m), \tag{3.9}$$

missä M on heikkojen oppijoiden lukumäärä, $H(x; \Theta_m)$ vastaa heikkoa oppijaa iteraatiossa m ja Θ_m vastaa heikkoon oppijaan H liittyviä parametreja [35, s. 356]. Parametrit Θ määrittävät koostepuun rakenteen. Jokaisella iteraatiolla tulee ratkaista estimaatit $\hat{\Theta}_m$ aiemman mallin f_{m-1} avulla

$$\hat{\Theta}_m = \arg\min_{\Theta_m} \sum_{i=1}^{N} L\left(y_i, f_{m-1}(x_i) + H(x_i; \Theta_m)\right), \tag{3.10}$$

missä L on valittu sakkofunktio. Suosittu sakkofunktio gradienttiboostauksessa on keskineliövirhe

$$L(y_i, f(x_i)) = \frac{1}{2}\left(y_i - f(x_i)\right)^2. \tag{3.11}$$

Nimensä mukaisesti gradienttiboostausmallin koulutuksessa käytetään gradientteja. Uusi heikko oppija koulutetaan laskemalla aiemman heikon oppijan sakkofunktion gradientti ja kulkemalla tätä gradienttia vastakkaiseen suuntaan, jolloin sakkofunktion arvo pienenee [35, s. 359]. Kun sakkofunktiona käytetään keskineliövirhettä 3.11 negatiiviseksi gradientiksi saadaan

$$-g_{im} = -\frac{\partial L\left(y_i, f(x_i)\right)}{\partial f(x_i)} = y_i - f(x_i), \tag{3.12}$$

mikä vastaa residuaalivirhettä [35, s. 360]. Koska gradientti on määritelty vain koulutukseen käytetyille havainnoille ja tavoitteena on yleistää malli myös uudelle aineistolle, koulutetaan uusi heikko oppija $H(x_i; \Theta_m)$, joka on mahdollisimman samansuuntainen negatiivisen gradientin $-g_{im}$ kanssa [35, s. 359]. Kun gradientin ja heikon oppijan etäisyyden mittana käytetään neliöityä virhettä, saadaan parametrien Θ_m estimaateiksi

$$\hat{\Theta}_m = \arg\min_{\Theta} \sum_{i=1}^{N} \left(-g_{im} - H(x_i; \Theta)\right)^2 \tag{3.13}$$

$$= \arg\min_{\Theta} \sum_{i=1}^{N} \left(y_i - f(x_i) - H(x_i; \Theta)\right)^2. \tag{3.14}$$

Kaavan 3.14 termit $y_i - f(x_i)$ vastaavat heikon oppijan H_{m-1} residuaaleja ja termi $H(x_i; \Theta)$ vastaa sovitettavaa heikkoa oppijaa H_m. Koska näiden termien erotus pyritään minimoimaan, uusi heikko oppija sovitetaan niin, että se minimoi edeltävän heikon oppijan residuaalivirheen.

Iteraation m heikko oppija voidaan kirjoittaa muodossa

$$f_m(\mathbf{x}) = f_{m-1}(\mathrm{x}) + H(\mathbf{x}; \hat{\Theta}_m). \tag{3.15}$$

Kun heikkoja oppijoita on sovitettu haluttu määrä M, saadaan vahvaksi oppijaksi

$$f_M(\mathbf{x}) = f_{M-1}(\mathbf{x}) + H(\mathbf{x}; \hat{\Theta}_M) \tag{3.16}$$

$$= \sum_{i=1}^{M} H(\mathbf{x}; \hat{\Theta}_i). \tag{3.17}$$

Gradienttiboostausmalleihin liittyviä hyperparametreja ovat esimerkiksi heikkojen oppijoiden koko ja määrä, joiden avulla gradienttiboostausmalleja voidaan säännellä [35, s. 364]. Kun heikkoina oppijoina käytetään päätöspuita, määrittyy niiden koko puiden syvyyden ja lehtien lukumäärän mukaan. Heikkojen oppijoiden lisääminen voidaan katsoa vastaavan epokkien lisäämistä neuroverkkojen tapauksessa. Aikaista lopetusta voidaan hyödyntää gradienttiboostauksessa niin, että heikkojen oppijoiden lisääminen lopetetaan silloin, kun yleistymisvirhe ei enää pienene [35, s. 364].

Gradienttiboostauksesta on tehty myös laskennallisesti tehokkaampia versioita, jotka sovelutvat erityisesti suurempien aineistojen mallintamiseen. Esimerkiksi Python-kirjastossa Scikit-Learn on implementoitu histogrammeihin perustuva gradienttiboostausmalli, jonka koulutus suurilla aineistolla on moninkertaisesti nopeampi kuin perinteinen gradienttiboostausmallin koulutus [36].

Myös gradienttiboostausmalleja voidaan täsmäkouluttaa uudella aineistolla. Tällöin mallissa säilyy jo koulutetut heikot oppijat, mutta malliin lisätään uusia heikkoja oppijoita, jotka minimoivat entisten heikkojen oppijoiden residuaalivirhettä uudessa aineistossa. Täsmäkoulutetun mallin tarkkuus on usein kuitenkin pienempi kuin mallin, jonka kouluttamiseen oltaisiin käytetty kerralla koko aineistoa [37].

3.3 Virhetarkastelu

Kun malli on sovitettu koulutusaineistoon, sen suorituskykyä voidaan tarkastella erilaisilla virhetermeillä. Usein regressiotehtävissä mallien suorituskykyä verrataan jäännösvirhehajonnalla (*root mean squared error*, RMSE)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \tag{3.18}$$

missä n on havaintojen lukumäärä, y_i on havainnon i todellinen arvo ja $\hat{y}_i$ on havainnon i mallinnettu arvo. Koska todellisen ja mallinnetun arvon erotus on korotettu toiseen

potenssiin, RMSE painottaa niiden ennusteiden arvoja, jotka ovat kauimpana todellisesta arvosta. Näitä kutsutaan vierashavainnoiksi (*outlier*) [31]. RMSE on myös erityisen käytetty PV-järjestelmien tehomallien vertailussa [8], [38]–[41]. Täydellisen mallin tapauksessa mallinnetut arvot vastaisivat täysin todellisia arvoja, jolloin $RMSE = 0$. Muussa tapauksessa RMSE on positiivinen arvo, ja paras malli RMSE:n perusteella on se, jolla on pienin arvo.

Jos aineistossa on paljon vierashavaintoja, voi olla mielekkäämpää käyttää virhetermiä, joka antaa vähemmän painoarvoa vierashavainnoille. Tällainen virhetermi on esimerkiksi keski-itseisvirhe (*mean absolute error*, MAE) [31]

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|. \tag{3.19}$$

Myös MAE on erityisen käytetty PV-järjestelmien tehomallien vertailussa [8], [38]–[41]. Myös MAE saa ei-negatiivisia arvoja (≥ 0), ja pienempi arvo tarkoittaa MAE:n mukaan parempaa mallia.

Edellä esitettyjen virhetermien lisäksi PV-järjestelmien tehomallien suorituskyvyn vertailuun voidaan käyttää selitysastetta R^2 (*coefficient of determination*) [8], [42]–[44]

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)}{\sum_{i=1}^{n}(y_i - \bar{y})}, \tag{3.20}$$

missä $\bar{y}$ on todellisten havaintojen y_i keskiarvo. R^2:lla voidaan tarkastella miten iso osa ulostulomuuttujan y vaihtelusta voidaan selittää käytetyn mallin avulla. R^2 on yleensä luku 0 ja 1 väliltä, mutta se voi joissain tapauksissa saada myös negatiivisia arvoja. Suurempi arvo tarkoittaa R^2:n mukaan parempaa mallia. [45]

Virhetermit RMSE ja MAE antavat arvon, josta ei voi yksinään päätellä mallin hyvyyttä, sillä virhetermien suuruus riippu käytetyn vastemuuttujan suuruudesta. Näiden virhetermien tulkinta voi olla vaikeaa ja niiden käyttö perustuukin eri malleista saatavien virhetermien suuruuksien vertailuun. R^2 antaa kuitenkin arvon, joka on helpompi tulkita, sillä sen arvo ei riipu vastemuuttujan suuruudesta. R^2 antamaa arvoa voidaan myös tulkita ilman, että sitä vertaillaan muihin malleihin. R^2:n käyttöä onkin suositeltu virhetarkasteluun regressiotehtävissä [45], ja tässäkin työssä keskitytään erityisesti R^2-arvojen tarkasteluun.

4. AINEISTO JA MENETELMÄT

4.1 Työssä käytetty aineisto

Työssä käytetty aineisto on peräisin neljästä eri sijainnista ja viidestä eri järjestelmästä. Havaintoja on Turun ammattikorkeakoulun Uuden energian ryhmän mittausasemalta Turusta (60,45°N; 22,30°E) ja Ilmatieteen laitoksen mittausasemilta Helsingistä (60,20°N; 24,96°E), Kuopiosta (62,89°N; 27,63°E) ja Sodankylästä (67,37°N; 26,65°E). Aineistot ovat usean vuoden pituisia.

Helsingin PV-järjestelmä on nimellisteholtaan käytetyistä PV-järjestelmistä suurin (21000 Wp). PV-järjestelmä sijaitsee kaupungissa Ilmatieteen laitoksen toimipisteen katolla. Paneelien atsiimuttikulma on 135 astetta pohjoisesta myötäpäivään. Paneelit on asennettu 15 asteen kallistuskulmaan. Tarkemmat järjestelmän paneeleihin liittyvät tiedot on esitetty taulukossa 4.1, josta löytyy myös muiden järjestelmien paneelien tiedot. Alhaisilla auringon tulokulmilla tapahtuu itseisvarjostamista (*self-shading*). Itseisvarjostaminen tarkoittaa sitä, että PV-paneelin varjo osuu sen takana olevaan paneeliin. Talvisin paneelien päälle kerääntyy lunta, sillä lumi ei pääse vapaasti putoamaan maahan paneelien päältä vaan kerääntyy helposti paneelien eteen katolle. Myös alhainen kallistuskulma helpottaa lumen kerääntymistä paneelin päälle. Järjestelmän tehon arvot näyttivät leikkaantuvan (*clipping*) eri kohdista joinakin vuosina, mikä voi johtua esimerkiksi invertterin toiminnan muutoksista. Koska leikkaantuminen voi vaikuttaa mallien toimintaan, koko aineiston tehon arvot rajoitetaan arvoon 20000 W ja tämän yli menevät havainnot poistetaan.

Kuopion PV-järjestelmä on nimelliseteholtaan käytetyistä PV-järjestelmistä toiseksi suurin (20280 Wp). PV-järjestelmä sijaitsee kaupungissa Ilmatieteen laitoksen toimipisteen katolla. Paneelien atsiimuttikulma on 217 astetta pohjoisesta myötäpäivään. Paneelit on asennettu 15 asteen kallistuskulmaan. Myös Kuopion järjestelmässä tapahtuu itseisvarjostamista. Helsingin järjestelmän tapaan Kuopion järjestelmässä paneelien päälle kerääntyy talvisin lunta. Järjestelmän tehon arvoissa näytti olevan selkeitä vierashavaintoja, jotka suodatettiin pois rajoittamalla tehon arvot arvoon 20280 W.

Sodankylän PV-järjestelmä koostuu kahdesta yhden paneelin järjestelmästä. Kummankin järjestelmän nimellisteho on 260 Wp. Toinen on asennettu 20 ja toinen 90 asteen kallistuskulmaan. Sodankylän PV-järjestelmä on 4 metrin korkuisella tasolla. Kummankin jär-

Taulukko 4.1. *PV-järjestelmien paneeleihin liittyvät parametrit.*

	Helsinki (IL)	Kuopio (IL)	Sodankylä (IL)	Turku (Turku AMK)
Leveysaste	60,20 °N	62,89 °N	67,37 °N	60,45 °N
Pituusaste	24,96 °E	27,63 °E	26,65 °E	22,30 °E
Paneelien lukumäärä	84	78	1 ja 1	18
Paneelien nimellisteho [Wp]	250	260	260	250
Järjestelmän nimellisteho [Wp]	21000	20280	260 ja 260	4500
Valmistaja	SolarWorld	SOLARWATT	SolarWorld	Kingdom Solar
Malli	SW 250 poly	BLUE 60P	SW 260 poly	KD-P250W
Atsimuuttikulma	135	217	115	180
Kallistuskulma	15	15	20 ja 90	15

jestelmän paneelin atsiimuttikulma on 115 astetta. Sodankylän järjestelmissä on päiviä, jolloin paneeleilla on lumipeitettä. Myös pystysuoralle paneelille kerääntyy välillä lumipeitettä. Paneelit on kuitenkin asennettu niin, että lumi putoaa sieltä maahan, eikä kasaannu tasolle paneelin eteen. Näin sulava lumi ei jää paneelin päälle.

Turun PV-järjestelmä (4500 Wp) poikkeaa muista järjestelmistä, sillä se on Turun ammattikorkeakoulun Uuden energian ryhmän ylläpitämä eikä Ilmatieteen laitoksen ylläpitämä. PV-järjestelmä sijaitsee kaupungissa rakennuksen katon päällä. Paneelit osoittavat etelään ja ovat 15 asteen kulmassa.

Aineistossa on havaintoja eri mittausresoluutioilla. Osa muuttujien mittauksista on suoritettu 1 minuutin välein (esimerkiksi teho ja säteilykomponentit), kun taas osa on suoritettu 10 minuutin (lähinnä paikallissää) tai 24 tunnin välein (lumipeite). Taulukossa 4.2 on esitetty eri mittauspisteiden mittausresoluutiot kullekin muuttujalle. Taulukkoon on merkitty myös tummalla muuttujat, joita tässä työssä käytetään. Käytetyt muuttujat muutetaan samaan resoluutioon imputoimalla puuttuvat arvot. Kymmenen minuutin resoluutiolla mitatut muuttujat imputoidaan korvaamalla edeltävät 9 arvoa samalla arvolla, joka on mitattu. 24 tunnin resoluutiolla mitattu lumen syvyys imputoidaan samalla tavalla, mutta nyt edelliset 1339 arvoa korvataan mitatulla arvolla. Lumen syvyys muuutetaan kaksiarvoiseksi muuttujaksi, joka ilmoittaa onko kyseisenä päivänä ollut lunta maassa vai ei. Sateen määrää ei käytetä tässä työssä, sillä Turun järjestelmän havaintopisteen mittaustapa poikkeaa muiden järjestelmien mittaustavasta. Kuten taulukosta 4.2 huomataan, Turun aineistossa ei ole lumen syvyyttä eikä pilvisyyttä. Lumen syvyys ja pilvisyys Turkuun on kuitenkin saatu Ilmatieteen laitoksen toimipisteeltä Artukaisista. Aineisto on ladattu havaintojen latauspalvelusta (https://www.ilmatieteenlaitos.fi/havaintojen-lataus). Aineisto on aikaväliltä 1.1.2018–21.10.2024 (lumen syvyys ladattu 21.10.2024 ja pilvisyys ladattu 5.12.2024). Turun havaintoasema mittaa vain GHI-säteilyä, joten POA-säteily tulee laskea kaavan 2.4 avulla. Tähän käytetään Python-kirjaston pvlib funktiota get_total_irradiance. Kaavan 2.4 tarvitsemat DNI- ja DHI-säteilyn arvot lasketaan GHI-säteilystä pvlib-kirjaston funktiolla pvlib.irradiance.erbs. Turun asemalla ei mitata myöskään paneelin lämpötilaa. Se laske-

Taulukko 4.2. *Muuttujien mittausresoluutiot mittauspisteillä.*

	Helsinki (IL)	Kuopio (IL)	Sodankylä (IL)	Turku (Turku AMK)
teho	**1 min**	**1 min**	**1 min**	**1 min**
GHI	**1 min**	**1 min**	**1 min**	**1 min**
DHI	1 min	1 min	1 min	-
DNI	1 min	1 min	1 min	-
POA	**1 min**	**1 min**	**1 min (90 astetta)**	**-**
ilman lämpötila	**10 min**	**10 min**	**10 min**	**1 min**
kosteus	**10 min**	**10 min**	**1 min**	**1 min**
pilvipeite	**10 min**	**10 min**	**1 min**	**-**
tuulen nopeus	**10 min**	**10 min**	**10 min**	**1 min**
tuulen suunta	**10 min**	**10 min**	**10 min**	**1 min**
paneelin lämpötila	**1 min**	**1 min**	**1 min**	**-**
katon lämpötila	1 min	1 min	-	-
ilmanpaine	10 min	10 min	1 min	-
alustan lämpötila	-	-	1 min	-
lumensyvyys	**24 h**	**24 h**	**24 h**	**-**

taan kaavan 2.11 avulla. Tähän käytetään pvlib-kirjaston funktiota sapm_module. Koska mikään asema ei mittaa kennon lämpötilaa, lasketaan se kaikille aineistoille kaavan 2.12 avulla. Kaavojen 2.11 ja 2.12 parametrit s_1, s_2 ja ΔT saadaan Python-kirjaston pvlib moduulista pvlib.temperature.TEMPERATURE_MODEL_PARAMETERS. Muuttujien laskeminen toisista muuttujista saattaa aiheuttaa aineistoon enemmän virhettä, kuin jos muuttujat olisi mitattu mittalaitteella. Lisäksi lumen syvyyttä ja pilvisyyttä ei olla mitattu Turussa PV-järjestelmän läheisyydessä. Tämän takia näiden muuttujien havainnot ladataan lähimmältä sääasemalta, mikä saattaa aiheuttaa aineistoon virhettä sillä esimerkiksi pilvisyys voi olla hyvin paikallista ja vaihdella nopeasti.

Taulukossa 4.2 esitettyjen muuttujien lisäksi koneoppimismallien koulutukseen käytetään muitakin muuttujia. Taulukossa 4.3 on esitetty kaikki työssä käytettävät muuttujat. Vasemmassa sarakkeessa on suoraan aineistosta saadut muuttujat, jotka on esitetty taulukossa 4.2 tummalla fontilla. Oikeassa sarakkeessa on järjestelmään liittyviä muuttujia sekä auringon sijaintiin liittyviä muuttujia.

Aineiston siistiminen on tärkeää, sillä puuttuvat ja virheelliset arvot voivat johtaa epätarkkoihin malleihin. Puuttuvien ja virheellisten arvojen huomioimisen lisäksi PV-data usein siistitään käyttämällä erilaisia suodattimia, joilla aineistosta saadaan yhteneväisempi ja soveliaampi haluttuun tarkoitukseen [46]. On tavallista käyttää esimerkiksi kirkkaan taivaan suodatusta (*clear-sky filter*), jotta esimerkiksi virhettä aiheuttava pilvisyys saadaan poistettua aineistosta. Tämä kuitenkin vähentää aineiston määrää, eikä tällaisella aineistolla koulutetut mallit välttämättä toimi hyvin pilvisellä kelillä. Jotta mahdollisimman pal-

Taulukko 4.3. *Koneoppimismallien kouluttamiseen käytetyt muuttujat.*

Mitatut muuttujat	Lisätyt muuttujat
teho	leveysaste
GHI	pituusaste
POA	korkeus
ilman lämpötila	paneelin kallistuskulma
kosteus	paneelin atsimuuttikulma
pilvisyys	auringon zeniittikulma
tuulen nopeus	auringon atsimuuttikulma
tuulen suunta	nimellisteho
paneelin lämpötila	paneelien ikä
lumensyvyys	kennon lämpötila
	vuodenaika

jon aineistoa pysyisi mukana ja luodut mallit pystyisivät estimoimaan myös pilvisempiä ja vaihtelevempia kelejä, tässä työssä ei käytetä kirkkaan taivaan suodatusta. Toistaiseksi ei ole olemassa laajasti käytettyjä ohjeita aineiston prosessointiin [47]. IEC esittää standardissa IEC 61724-1 [12] yleisiä suosituksia aineiston käsittelyyn. Standardi ei kuitenkaan esitä tarkempia tutkimuskohtaisia suosituksia aineiston prosessointiin. Standardin pohjalta on kuitenkin tehty yksityiskohtaisempia tutkimuksia aineiston käsittelyyn [47]. Tässä työssä aineiston prosessointiin sovelletaan Livera ym. tutkimuksen [47] suosituksia, jotka pohjautuvat IEC 61724 standardiin sekä muihin aikaisempiin töihin.

Aluksi aineistosta poistetaan ne havainnot, joissa säteilyä on vähän. Tällä poistetaan esimerkiksi yöhavainnot pois aineistosta. Rajana voidaan käyttää esimerkiksi POA < 20 W/m² [47]. Koska tässä työssä halutaan luoda malli, joka toimii myös pilvisillä ja lumisilla keleillä, käytetään säteilyn suodatuksen rajana POA < 5 W/m². Tällä saadaan pidettyä useampi havainto aineistossa, mutta yöt suodattuvat tällä kuitenkin pois. Aineistosta suodatetaan pois myös havainnot, joissa toteutunut teho on alle kaksi prosenttia järjestelmän ilmoitetusta nimellistehosta. Suodatusta kokeiltiin myös yhden prosentin raja-arvolla, mutta kuvaajista päätellen aineistoon jäi suodatuksen jälkeen vielä huomattavasti vierashavaintoja, joissa on mitattu säteilyä, muttei tehoa. Myös kahden prosentin suodatuksen jälkeen kuvaajista on havaittavissa tällaisia vierashavaintoja, mutta selkeästi vähemmän kuin yhden prosentin suodatuksella. Sitten aineistosta etsitään virheelliset arvot. Niitä voidaan etsiä erilaisten kynnysarvojen avulla. Livera ym. suosittelevat tutkimuksessa [47] käyttämään kynnysarvoja mitatuille arvoille. Havainnot, joissa jokin muuttuja on kynnysarvojen ulkopuolella, poistetaan aineistosta. Tässä työssä käytetyt kynnysarvot on esitetty taulukossa 4.4. Niiden pohjana on käytetty tutkimuksessa [47] suositeltuja kynnysarvoja.

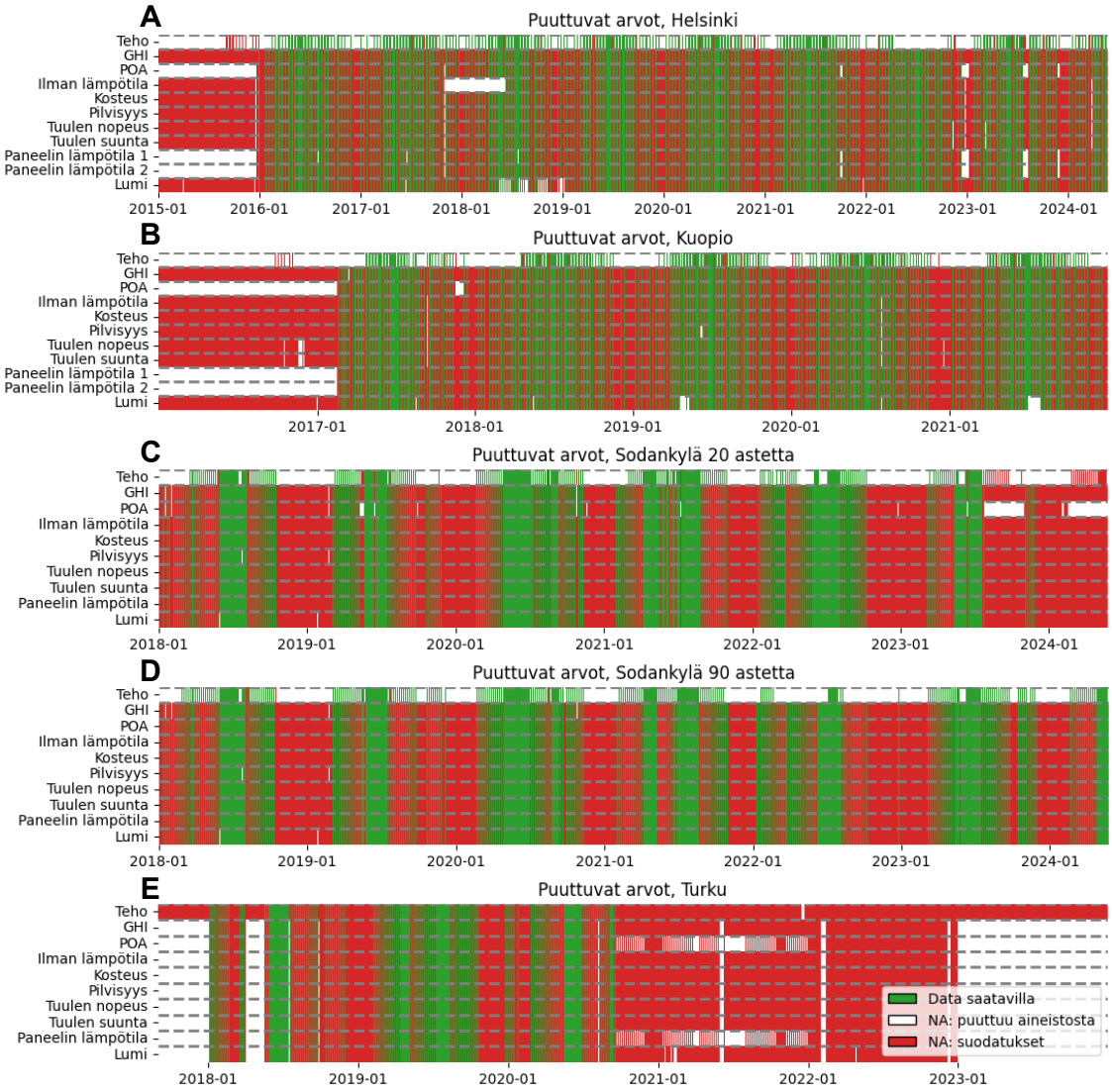
Kuvassa 4.1 on esitetty aineistojen puuttuvat arvot niille muuttujille, joita työssä käytetään.

Taulukko 4.4. *Käytetyt kynnysarvot aineistolle. Muokattu lähteestä [47].*

Muuttuja	Minimi	Maksimi
G_{POA} [W/m²]	5	1300
$G_{GHI}, G_{DHI}, G_{DNI}$ [W/m²]	0	$-$
Ilman lämpötila [℃]	-40	60
Paneelin lämpötila [℃]	-40	120
Tuulen nopeus [m/s]	0	32
Teho [W]	$2\ \%$ nimellistehosta	$-$

Kuvassa näkyy pidemmät puuttuvat ajanjaksot, mutta yksittäisiä puuttuvia havaintoja ei huomaa kuvasta. Kuvista 4.1A, 4.1B, 4.1C ja 4.1D huomataan, että erityisesti tehossa on ajanjaksoja, jotka puuttuvat kokonaan aineistosta. Puuttuvien havaintojen syytä ei voida varmuudella päätellä pelkästään aineiston perusteella, mutta kyseessä voi olla esimerkiksi invertterin vikaantuminen. Nämä ajanjaksot on esitetty valkoisella. Taulukon 4.4 mukaisen suodatuksen takia poistuneet havainnot on merkitty punaisella. Huomataan, että kaikissa aineistoissa vaikuttaa olevan suodattuneita (punaisella merkittyjä) arvoja tasaisin väliajoin. Erityisesti talvisin aineistoissa näyttäisi olevan pidempiä puuttuvia ajanjaksoja. Tämä on toisaalta odotettu tulos ja johtuu esimerkiksi talven vähäisemmästä säteilystä ja lumipeitteestä paneelin päällä. Turun järjestelmän kuvassa 4.1E ei ole niin paljon puuttuvia tehon arvoja, kuin muissa aineistoissa, mutta aineisto suodattuu silti pois vuodesta 2021 eteenpäin. Puuttuvien arvojen käsittelyyn on erilaisia menetelmiä [47]. Tässä työssä aineistosta poistetaan ne havainnot, joissa on yksikin puuttuva arvo.

Koska osaa fysikaalisista malleista voidaan käyttää suoraan ilman kouluttamista, niitä voidaan pitää yleisinä malleina, joita voidaan käyttää suoraan eri PV-järjestelmissä ilman aiempaa aineistoa. Tämä johtaa kysymykseen voidaanko koneoppimismallien avulla tehdä yleinen tehomalli, joka sopii Suomen olosuhteisiin fysikaalisia malleja paremmin. Siksi koulutettuja koneoppimismalleja olisi hyvä testata PV-järjestelmällä, joka on täysin uusi koulutetuille malleille. Tämän takia yksi järjestelmistä jätetään mallien testausta varten eikä sitä käytetä mallien kouluttamiseen. Testaukseen valitaan Kuopion järjestelmä, jotta mallien kouluttamiseen saadaan aineistoa sekä etelästä että pohjoisesta. Koska Kuopio on keskellä Suomea ja kaukana koulutukseen käytettävistä järjestelmistä, voivat Kuopion sääolosuhteet poiketa mallien kouluttamiseen käytetyistä sääolosuhteista. Näin vältytään siltä, että mallit antaisivat testauksessa liian optimistisia tuloksia sen takia, että testattava järjestelmä muistuttaa sääolosuhteiltaan koulutuksessa käytettyä järjestelmää. Helsingin, Turun ja Sodankylän järjestelmien aineistot yhdistetään koulutusaineistoksi. Koulutusaineistosta jätetään kuitenkin vuosi 2019 pois, jotta mallien suorituskykyä voidaan tarkastella myös malleille tutuilla järjestelmillä. Fysikaalisia malleja 6k ja PVUSA voidaan myös kouluttaa tiettyyn järjestelmään sopiviksi sovittamalla niiden kertoimet uudestaan. Tämä vuoksi myös koneoppimismallit täsmäkoulutetaan Kuopion aineistoon, jotta nähdään op-



Kuva 4.1. *Aineistojen puuttuvat arvot. Kuvassa A on esitetty Helsingin järjestelmä, kuvassa B Kuopion järjestelmä, kuvassa C Sodankylän järjestelmä, jossa paneeli on asennettu 20 asteen kulmaan, kuvassa D Sodankylän järjestelmä, jossa paneeli on asenettu 90 asteen kulmaan ja kuvassa E Turun järjestelmä.*

pivatko yleiset koneoppimismallit paikallisen järjestelmän ominaispiirteitä ja parantuuko mallien suorituskyky tämän avulla.

Ennen kuin mallit koulutetaan, aineistot skaalataan. Koulutusaineiston numeeriset muuttujat GHI-säteily, POA-säteily, ilman lämpötila, kosteus, tuulen nopeus, tuulen suunta, paneelin lämpötila, leveysaste, pituusaste, korkeus, paneelin kallistuskulma, paneelin atsimuuttikulma, auringon zeniittikulma, auringon atsimuuttikulma, nimellisteho, paneelien ikä ja kennon lämpötila normalisoidaan Python-kirjaston Scikit-Learn luokalla sklearn.preprocessing.StandardScaler. Muuttujat vuodenaika, pilvisyys ja lumensyvyys muutetaan yksi-kuuma-muotoon Scikit-Learnin luokalla sklearn.preprocessing.OneHotEncoder. Testiaineisto ja vuoden 2019 aineisto skaalataan samalla tavalla kuin koulutusaineisto. Te-

ho skaalataan jakamalla se järjestelmän nimellisteholla. Tehon jakaminen nimellisteholla skaalaa kunkin järjestelmän tehon samaan suuruusluokkaan. Tehon arvot ovat skaalauksen jälkeen yleensä 0 ja 1 välillä. On kuitenkin mahdollista, että tuotettu teho ylittää nimellistehon esimerkiksi pilviheijastuksen (*cloud enhancement*) aikana, jolloin pilvistä heijastuva säteily voi lisätä paneelille osuvaa säteilyä jopa 50 % kirkkaan taivaan säteilyyn verrattuna [48]. Nimellisteholla skaalattu arvo voi tällöin ylittää arvon 1.

4.2 Fysikaalisten mallien koulutus

Työssä käytettävät fysikaaliset mallit 6k 2.13 ja PVUSA 2.15 ovat regressiomalleja, joiden kertoimet pitää sovittaa kouluttamalla ne osalla aineistolla. Kertoimet on kuitenkin laskettu valmiiksi 6k-mallin eri paneelimateriaaleille ja asennuksille. Nämä valmiit kertoimet saadaan suoraan Python-kirjastosta pvlib, kun paneelien materiaali ja asennus tiedetään. Koska valmiiksi lasketut kertoimet eivät välttämättä kuvaa työssä käytettyjä järjestelmiä parhaiten, sovitetaan mallien kertoimet myös jokaiseen järjestelmään. PVUSA-malli sovitetaan suoraan käytettyyn aineistoon, sillä sille ei ole laskettu valmiita kertoimia ainakaan työssä käytetyissä Python-kirjastoissa. Fysikaalinen malli PVWatts 2.15 ei tarvitse sovittamista, koska siinä ei ole kertoimia.

6k- ja PVUSA-malli koulutetaan Python kirjaston SciPy funktiolla scipy.optimize.curve_fit. 6k-mallin kertoimien alkuarvoina käytetään julkaisussa [5] saatuja kertoimia Helsingin järjestelmälle. PVUSA-mallin kertoimien alkuarvoina käytetään julkaisussa [49] saatuja kertoimia. Kertoimet sovitetaan jokaiselle järjestelmälle erikseen käyttämällä kunkin järjestelmän aineiston 365 ensimmäistä päivää. Mallien kertoimien alkuarvot ja sovitetut kertoimet on esitetty taulukossa 4.5. 6k-mallin kertoimet on skaalattu järjestelmän nimellisteholla kaavan 2.14 mukaisesti, jotta eri järjestelmien kertoimia on helpompi verrata keskenään. 6k- ja PVUSA-mallien sovitetut kertoimet eroavat alkuarvoista sekä suuruudeltaan että välillä myös etumerkiltään. Koska 6k- ja PVUSA-malleissa on muuttujien välisiä yhdysvaikutustermejä (kuten esimerkiksi 6k-mallin kolmatta kerrointa vastaava POA-säteilyn ja paneelin lämpötilan yhdysvaikutustermi $G'_{POA} \cdot T'$), on kertoimien tulkinta haastavaa.

4.3 Koneoppimismallien koulutus

Työssä käytetään kahta eri koneoppimismallia. Ensiksi aineistoon sovitetaan MLP-malli ja sitten HGBR-malli. Kumpikin malli muodostetaan Pythonin Scikit-Learn-kirjaston avulla. MLP-malli muodostetaan luokan sklearn.neural_network.MLPRegressor avulla. HGBR-malli muodostetaan luokan sklearn.ensemble.HistGradientBoostingRegressor avulla. Scikit-Learn-kirjasto suosittelee histogrammeihin perustuvan gradienttiboostausmallin käyttöä, kun aineisto on sen kokoinen, että siinä on kymmeniä tuhansia havaintoja [36]. Tässä työssä koulutusaineisto sisältää 2404641 havaintoa, jolloin histogrammipohjaisen mallin

Taulukko 4.5. *Fysikaalisten mallien kertoimet. 6k-mallin kertoimien alkuarvoina käytetään julkaisussa [5] saatuja kertoimia Helsingin järjestelmälle ja saadut kertoimet on skaalattu nimellisteholla. PVUSA-mallin kertoimien alkuarvoina käytetään julkaisussa [49] saatuja kertoimia.*

Kertoimet		Alkuarvot	Helsinki	Kuopio	Sodankylä 20 astetta	Sodankylä 90 astetta	Turku
6k	k_1	$0,045866$	$0,220571$	$0,212617$	$0,475013$	$0,293376$	$0,261100$
	k_2	$-0,002035$	$0,089057$	$0,084644$	$0,275074$	$0,081564$	$0,083557$
	k_3	$-0,006095$	$0,004960$	$-0,003490$	$-0,000255$	$-0,003798$	$-0,011361$
	k_4	$-0,000120$	$0,000973$	$-0,004579$	$0,009496$	$0,001019$	$-0,019583$
	k_5	$0,000372$	$-0,001321$	$-0,000067$	$0,009079$	$0,000808$	$-0,005910$
	k_6	$0,000004$	$-0,000644$	$-0,000109$	$-0,000243$	$-0,000167$	$0,000038$
PVUSA	a	$1,41870$	$14,4231$	$17,2039$	$0,217539$	$0,204210$	$3,89947$
	b	$0,000051$	$0,001288$	$0,000895$	$-0,000046$	$0,000017$	$-0,000481$
	c	$0,002291$	$0,107726$	$0,092785$	$0,001804$	$0,001009$	$0,054009$
	d	$0,000361$	$0,159385$	$-0,008405$	$0,000668$	$-0,000178$	$0,001951$

Taulukko 4.6. *MLP-mallin koulutukseen kokeillut hyperparametrit.*

Hyperparametri	Kokeillut arvot
piilokerrokset	$(5,5,5),\ (10,10,10),\ (20,20,20),\ (50,50,50),\ (100,100,100)$
oppimisnopeuden alustus	$1 \cdot 10^{-2},\ 1 \cdot 10^{-3},\ 1 \cdot 10^{-4}$
oppimisnopeuden muutos	adaptive
piilokerrosten aktivaatiofunktio	relu
eräkoko	32
algoritmi	adam

kouluttaminen on todennäköisesti paljon nopeampaa kuin perinteisen gradienttiboostausmallin kouluttaminen.

Koneoppimismallien hyperparametrit valitaan ruutuetsinnällä, joka on toteutettu Scikit-Learn-kirjaston luokassa sklearn.model_selection.GridSearchCV. MLP-mallin kokeillut hyperparametrit on esitetty taulukossa 4.6. Vaikka mallin valinnassa käytetään CSC:n supertietokonetta Puhti, on MLP-mallin koulutus laskennallisesti raskasta ja hidasta, joten kokeiltavien hyperparametrien lukumäärä on pidetty maltillisena. Koska ruutuetsintä laskee jokaisen annetun hyperparametrien yhdistelmän, uuden kokeiltavan hyperparametrin lisäys kasvattaa mallin valinnassa vertailtavien mallien määrää nopeasti. Parhaaksi MLP-malliksi valikoituu malli, jossa piilokerrosten leveys on 10 neuronia ja jonka oppimisnopeus on alustettu arvoon $1 \cdot 10^{-3}$. HGBR-mallin kokeillut hyperparametrit on esitetty taulukossa 4.7. Parhaaksi HGBR-malliksi valikoituu malli, jossa oppimisnopeus on $1 \cdot 10^{-2}$, heikkojen oppijoiden maksimimäärä on 6400 ja heikkojen oppijoiden lehtien maksimimäärä on 64. Vaikka heikkojen oppijoiden maksimimäärä on 6400, niitä sovitetaan malliin vain 3601 kappaletta, sillä tällöin mallin validaatiovirhe on pienin ja aikainen lopetus lopettaa uusien heikkojen oppijoiden sovituksen.

Kun koneoppimismallit on koulutettu, kokeillaan niitä aluksi suoraan Kuopion aineistoon.

Taulukko 4.7. *HGBR-mallin koulutukseen kokeillut hyperparametrit.*

Hyperparametri	Kokeillut arvot
heikkojen oppijoiden maksimimäärä	$200,\ 400,\ 800,\ 1600,\ 3200,\ 6400$
lehtien maksimimäärä	$16,\ 32,\ 64,\ 128$
oppimisnopeus	$2 \cdot 10^{-1},\ 1 \cdot 10^{-1},\ 1 \cdot 10^{-2},\ 1 \cdot 10^{-3}$

Tällöin voidaan tarkastella miten hyvin mallit toimivat yleisinä tehomalleina niille täysin uudessa järjestelmässä. Tämän lisäksi mallit täsmäkoulutetaan Kuopion aineistoon. Täsmäkoulutukseen käytetään 365 ensimmäisen päivän aineistoa Kuopion aineistosta, kuten fysikaalisten mallien täsmäkoulutuksessa. Koneoppimismallien täsmäkoulutus tapahtuu käyttämällä MLPRegressor- ja HistGradienBoostingRegressor-luokissa parametria warm_start. Tällöin MLP-malli jatkaa painojen optimointia Kuopion aineistoon. Painojen alkuarvoina käytetään niitä painojen arvoja, joihin ennen täsmäkoulutusta päästiin. HGBR-mallin tapauksessa aikaisempia heikkoja oppijoita ei muokata, vaan Kuopion aineiston pohjalta malliin sovitetaan uusia heikkoja oppijoita [50]. Täsmäkoulutuksessa on myös käytössä aikainen lopetus, mikä tarkoittaa, että koulutus lopetetaan mikäli validaatiovirhe ei vähene. Koska HGBR-mallin validaatiovirhe nousee heti ensimmäisen uuden heikon oppijan sovituksen jälkeen, täsmäkoulutus lopetetaan. Täsmäkoulutettu HGBR-malli ei siis poikkea paljoakaan alkuperäisestä mallista, sillä siinä on vain yksi uusi heikko oppija.

5. TULOKSET

Työssä käytetään aineistoa viidestä eri järjestelmästä. Osa fysikaalisista malleista ei tarvitse erillistä kouluttamista, jotta niitä voidaan käyttää tehon mallintamiseen. Tällainen on tässsä työssä käytetty PVWatts-malli. Koska 6k-mallille on laskettu oletuskertoimet, voi myös 6k-mallia käyttää suoraan järjestelmän tehon mallintamiseen ilman koulutusta. Toisaalta PVUSA-malli tarvitsee kouluttamista kertoimien sovitukseen ja 6k-mallin kertoimet voi halutessaan sovittaa itse. Koneoppimismallit on sen sijaan koulutettu useamman järjestelmän aineistolla ja niitä testataan Kuopion järjestelmässä. Jotta tulokset fysikaalisten ja koneoppimismallien välillä olisivat mahdollisimman helposti vertailtavissa, on myös fysikaalisten mallien kohdalla rajoituttu tarkastelemaan fysikaalisten mallien toimintaa erityisesti Kuopion järjestelmässä, vaikka fysikaalisia malleja voidaankin käyttää jokaisessa järjestelmässä. Malleihin, joita ei ole erikseen sovitettu käytettyyn järjestelmään, viitataan nimityksellä yleinen malli. Malleihin, jotka on erikseen sovitettu käytettyyn järjestelmään, viitataan nimityksellä täsmäkoulutettu malli.

Kuvassa 5.1 on esitetty kunkin järjestelmän tuottama teho ja POA-säteily aineiston suodatuksen jälkeen. Koska teholla ja POA-säteilyllä on vahva korrelaatio, tulisi kuvaajien pisteiden asettua hyvin suoralle. Kuvaajiin on merkattu vihreällä ne havainnot, jolloin maassa ei ole lunta ja punaisella ne havainnot, jolloin maassa on mitattu lunta. Vaikka maassa olisi lunta, se ei kuitenkaan välttämättä tarkoita, että lunta olisi paneelien päällä. Ja vaikka aineisto näyttäisi, ettei maassa ole lunta, on päivän aikana voinut sataa lunta, mikä ei näy aineistossa, koska lumensyvyys mitataan vain kerran päivässä. Kuvista huomataan, että lumi aiheuttaa joissain järjestelmissä poikkeavia arvoja. Erityisesti Helsingin järjestelmän kuvaajassa 5.1A lumi näyttäisi haittaavan tehon tuotantoa, sillä osalla havainnoista, joissa maassa on lunta, on alhaisempi teho mitä POA-säteilyn mukaan olisi odotettavissa. Tällaiset poikkeavat arvot voivat johtua siitä, että paneelin päällä on lunta, mutta säteilyä mittaavan mittalaitteen päällä ei. Tällöin mittalaite raportoi säteilyä auringon paistaessa, mutta paneelit eivät tuota tehoa lumikerroksen takia. Myös Kuopion ja Sodankylän 20 asteen järjestelmän kuvaajissa 5.1B ja 5.1C lumella näyttäisi olevan välillä tehon tuotantoa haittaava vaikutus. Toisaalta Sodankylän 90 asteen järjestelmän kuvaajassa 5.1D lumella näyttäisi olevan sekä säteilyä että tehon tuotantoa vahvistava vaikutus. Tämä johtuu todennäköisesti siitä, että pystysuoraan asennetun paneelin päälle ei kerääny niin paljoa lunta. Myös maassa oleva lumi heijastuu paremmin pystysuoralle paneelille, jolloin sekä

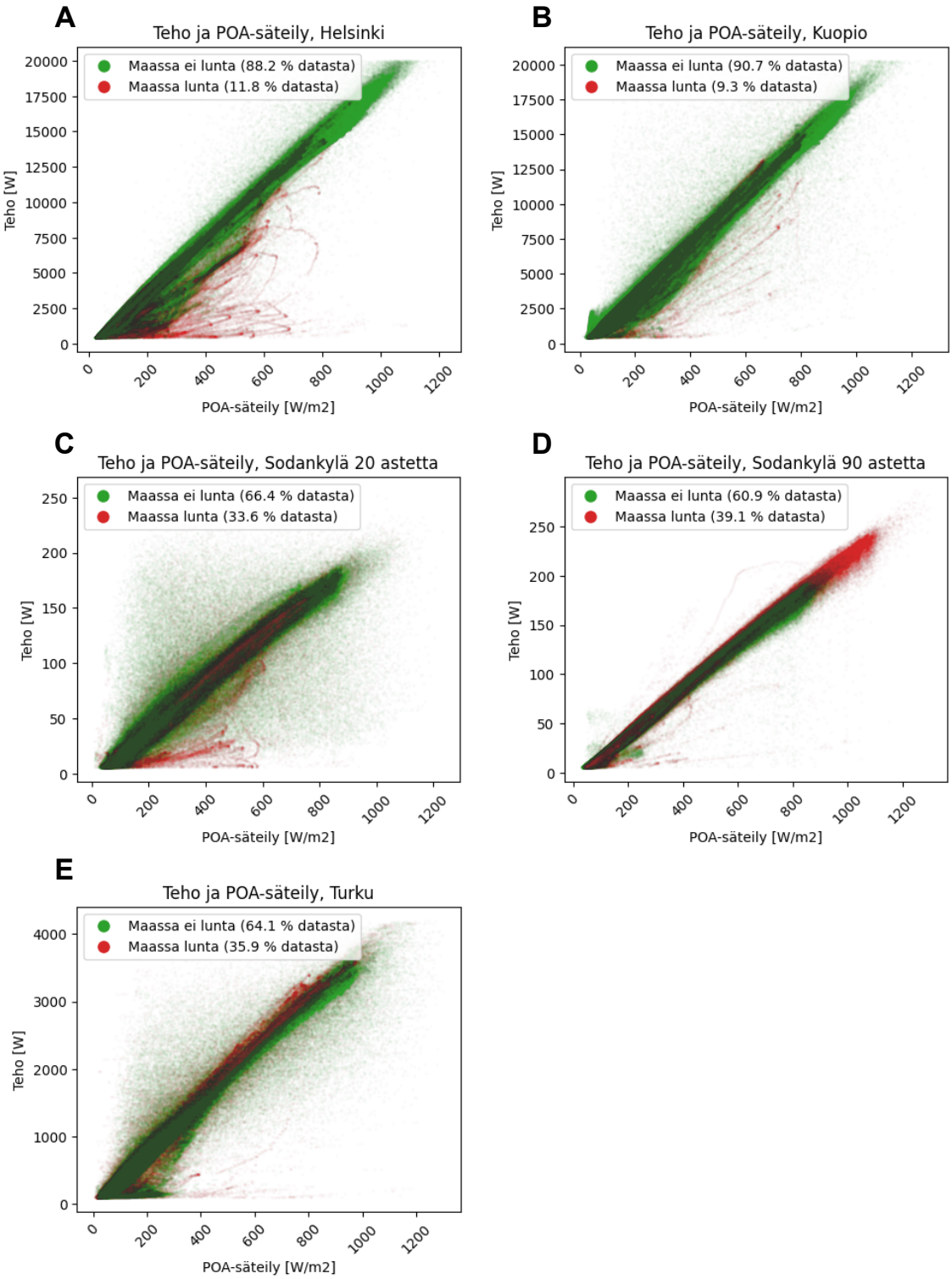
POA-säteily että tehon tuotanto tehostuvat. Turun järjestelmän kuvaajassa 5.1E lumiset havainnot näyttäisivät osuvan suoralle eikä poikkeavia havaintoja juurikaan ole. On kuitenkin yllättävää, että Turun suodatetussa aineistossa noin 36 % havainnoista sisältää lunta. Tämä voi johtua kuitenkin esimerkiksi siitä, että osa kesän havainnoista on puuttuvien arvojen takia suodattunut pois, jolloin lumisten havaintojen osuus koko aineistosta kasvaa. Kuvaajista 5.1C ja 5.1E nähdään, että Sodankylän 20 asteen ja Turun järjestelmissä on muita järjestelmiä enemmän kohinaa. Tämä voi johtua monesta syystä, kuten virheellisistä mittauksista. Koska tässä työssä aineistoa suodatetaan vain taulukon 4.4 kynnysarvojen mukaan, aineistojen kohinaa ei suodateta pois.

5.1 Fysikaalisten mallien tulokset

Fysikaaliset mallit on muodostettu kappaleessa 4.2 esitetyllä tavalla. Taulukossa 5.1 on esitetty fysikaalisten mallien R^2-arvot sekä MAE- ja RMSE-virhetermit eri järjestelmille. Tulosten perusteella mallien täsmäkouluttaminen kasvattaa R^2-arvoa ja pienentää MAE- ja RMSE-virhetermien arvoja, mikä on odotettu tulos. Yleiset 6k- ja PVWatts-mallit näyttäisivät virhetermien perusteella toimivan lähes yhtä hyvin käytetyissä järjestelmissä. Myös täsmäkoulutetut 6k- ja PVUSA-mallit näyttäisivät virhetermien mukaan toimivan lähes yhtä hyvin. Tulos poikkeaa tutkimuksen [6] tuloksista, joissa täsmäkoulutettu PVUSA-malli toimi selvästi täsmäkoulutettua 6k-mallia huonommin.

R^2-arvot ovat Sodankylän 20 asteen järjestelmää lukuun ottamatta aika suuria erityisesti täsmäkoulutetuilla malleilla. Sodankylän 20 asteen suuret virhetermit ja pienet R^2-arvot voivat johtua esimerkiksi aineiston suuresta kohinasta, joka näkyy kuvassa 5.1C. Mutta esimerkiksi täsmäkoulutetun 6k-mallin R^2-arvo Helsingissä on 0,97, mikä viittaa mallin hyvään suorituskykyyn. Koska pohjoisista olosuhteista ei löytynyt tutkimuksia, joissa tarkastellaan tehomalleja R^2-termin avulla, on tulosta verrattava tutkimuksiin muista sijainneista. Esimerkiksi koneoppimispohjaisille tehomalleille on raportoitu niinkin isoja R^2-arvoja kuin 0,998 (Algeria) [8], 0,993 (Tansania) [42] ja 0,99999 (Espanja) [44]. Verrattuna näihin tuloksiin R^2-arvo 0,97 ei välttämättä vaikuta kovin hyvältä. Koska pohjoiset olosuhteet tekevät tehon mallintamisesta hankalaa ja koska tässä työssä aineistoa on suodatettu tarkoituksella vähän, jotta aineistoon jää myös esimerkiksi haasteita aiheuttavia lumihavaintoja, voidaan R^2-arvo 0,97 tulkita hyväksi. Joissain tutkimuksissa on myös raportoitu pienempiä R^2-arvoja, kuten 0,87 (Intia) [43], mikä viittaisi myös siihen, että esimerkiksi R^2-arvo 0,97 voidaan tulkita hyväksi.

Böök ym. tarkasteli tutkimuksessa [5] samoja Helsingin ja Kuopion järjestelmiä kuin tässä työssä. Vaikka he eivät tarkastelleet R^2-arvoja, he laskivat MAE- ja RMSE-virhetermit mallinnetulle teholle (tutkimuksen [5] taulukko 7). He jakoivat virhetarkastelun lumettomille keleille ja lumisille keleille. Heidän tutkimuksessa teho on skaalattu yksikköön W/kWP, joten virhetermit on jaettava arvolla 1000, jotta ne vastaavat tämän työn virhetermejä.

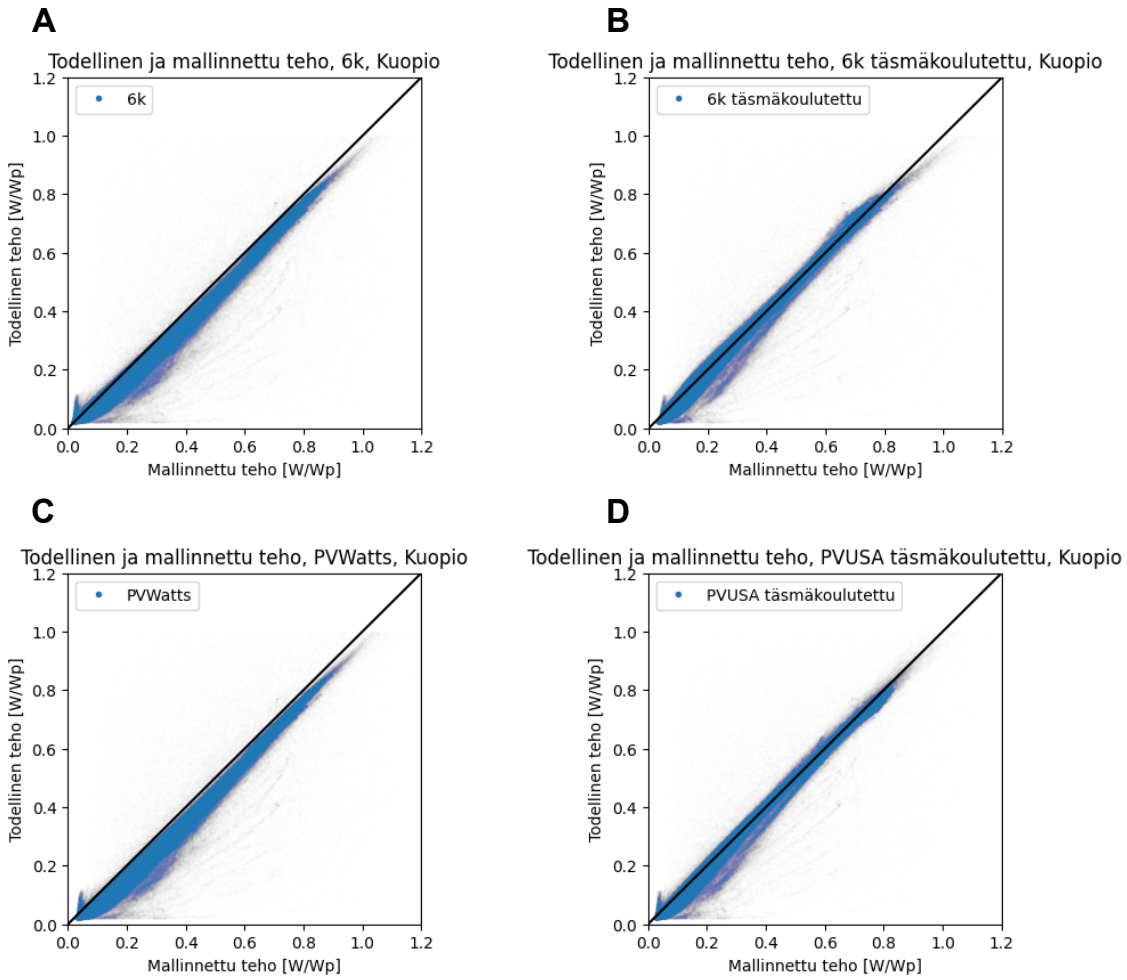


Kuva 5.1. *Lumen vaikutus mitattuun POA-säteilyyn ja tehoon.*

Taulukko 5.1. *Fysikaalisten mallien virhetermit. Yleisen 6k-mallin ja PVWatts-mallin virhetermit lasketaan suoraan järjestelmän aineistoon. Täsmäkoulutetun 6k-mallin ja PVUSA-mallin kouluttamiseen käytetään kunkin järjestelmän aineiston ensimmäiset 365 päivää, ja virhetermit lasketaan jäljelle jääneeseen aineistoon.*

Järjestelmä	Malli	R^2	MAE	RMSE
Helsinki (IL) 2015–2024	6k (yleinen)	$0,91$	$0,05$	$0,07$
	6k (täsmäkoulutettu)	$0,97$	$0,03$	$0,04$
	PVWatts (yleinen)	$0,90$	$0,05$	$0,08$
	PVUSA (täsmäkoulutettu)	$0,97$	$0,03$	$0,05$
Kuopio (IL) 2017–2021	6k (yleinen)	$0,93$	$0,04$	$0,06$
	6k (täsmäkoulutettu)	$0,96$	$0,03$	$0,05$
	PVWatts (yleinen)	$0,92$	$0,04$	$0,07$
	PVUSA (täsmäkoulutettu)	$0,96$	$0,03$	$0,05$
Sodankylä 20 astetta (IL) 2018–2024	6k (yleinen)	$0,56$	$0,09$	$0,13$
	6k (täsmäkoulutettu)	$0,76$	$0,06$	$0,09$
	PVWatts (yleinen)	$0,56$	$0,09$	$0,13$
	PVUSA (täsmäkoulutettu)	$0,82$	$0,05$	$0,08$
Sodankylä 90 astetta (IL) 2018–2024	6k (yleinen)	$0,89$	$0,06$	$0,08$
	6k (täsmäkoulutettu)	$0,96$	$0,03$	$0,05$
	PVWatts (yleinen)	$0,90$	$0,07$	$0,08$
	PVUSA (täsmäkoulutettu)	$0,98$	$0,02$	$0,03$
Turku (Turku AMK) 2018–2020	6k (yleinen)	$0,85$	$0,06$	$0,09$
	6k (täsmäkoulutettu)	$0,92$	$0,04$	$0,07$
	PVWatts (yleinen)	$0,85$	$0,07$	$0,09$
	PVUSA (täsmäkoulutettu)	$0,93$	$0,04$	$0,06$

Tutkimuksen [5] mukaan virhetermit ovat huomattavasti suurempia lumisilla kuin lumettomilla keleillä. Esimerkiksi täsmäkoulutetun 6k-mallin MSE-virhe lumettomalla kelillä on 0,017, kun taas lumisella kelillä se on 0,117 (MSE-arvot on jaettu arvolla 1000, jotta niitä voidaan verrata tämän työn virhetermeihin). Täsmäkoulutettu 6k-malli antoi tässä työssä Kuopion järjestelmälle MSE-virhetermin arvoksi 0,03. Koska arvo pitää sisällään lumettomat ja lumiset kelit, on odotettavaa, että tämän työn MSE-virhetermi on suurempi kuin tutkimuksen [5] MSE-virhetermi lumettomille keleille. Kun tämä otetaan huomioon, vaikuttaa tämän työn tulokset olevan linjassa tutkimuksen [5] kanssa. Myös täsmäkoulutetun 6k-mallin RMSE-virhetermi antoi tässä työssä lukuja, jotka sijoittuivat tutkimuksen [5] lumettoman ja lumellisen RMSE-virhetermien väliin.



Kuva 5.2. *Fysikaalisten mallien todellisten ja mallinnettujen tehojen pistekuvaajat Kuopion järjestelmässä. Täsmäkoulutetun 6k-mallin ja PVUSA-mallin kouluttamiseen käytetään Kuopion järjestelmän aineiston ensimmäiset 365 päivää.*

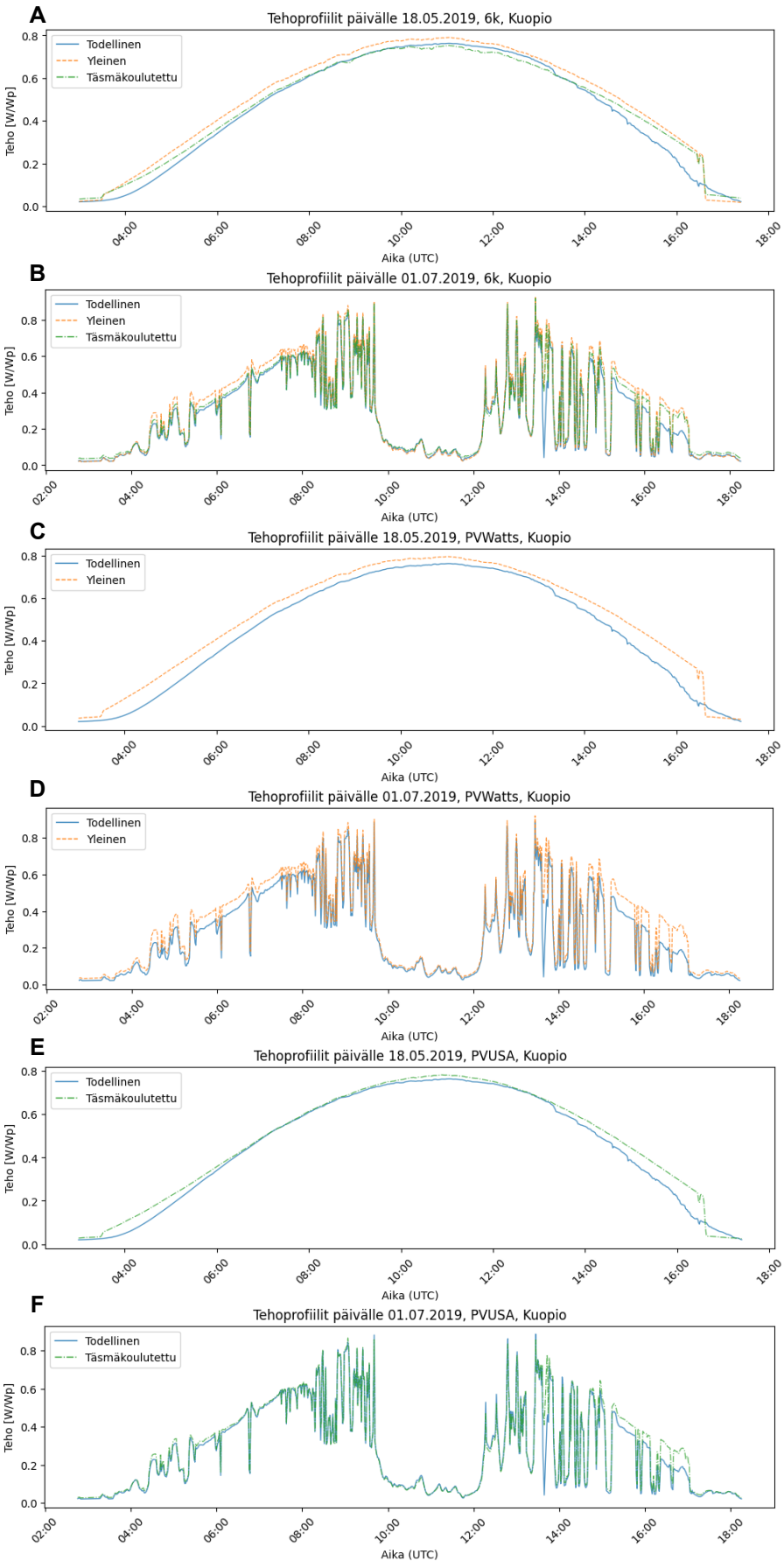
Kuvassa 5.2 on esitetty eri malleille pistekuvaajat, joista näkyy miten hyvin mallinnetut tehot vastaavat todellisia tehon arvoja Kuopion järjestelmässä. Täydellisen mallin tapauksessa pisteet osuisivat mustalle suoralle. Kuvista huomataan, että yleiset 6k- ja PVWatts-mallit mallintavat lähes kaikissa tilanteissa tehon liian suureksi, sillä pisteet ovat mustan suoran alapuolella. Tämä voisi johtua esimerkiksi järjestelmän varjostumisesta, jolloin järjestelmä ei tuotakaan niin paljon tehoa, kuin säteilyn perusteella olisi odotettavissa. Vaikka pisteet ovat lähes systemaattisesti mustan suoran alapuolella, pisteparvi on kuitenkin oikean muotoinen eikä siinä näytä olevan kovin suurta hajontaa. Täsmäkoulutetut 6k- ja PVUSA-mallit istuvat suoralle paremmin, mutta alhaisilla tehon arvoilla nämäkin mallit mallintavat tehon liian suureksi lievästä kaaresta päätellen todellisen tehon arvojen 0,0 ja 0,3 välillä.

Kuvassa 5.3 on esitetty kaksi eri päivää, joiden avulla verrataan mallien mallintamaa tehoprofiilia toteutuneeseen tehoprofiiliin. Ensin tarkastellaan aurinkoista päivää, jonka tehoprofiili on tasainen ja kellokäyrän muotoinen. Tämän jälkeen tarkastellaan pilvistä päivää,

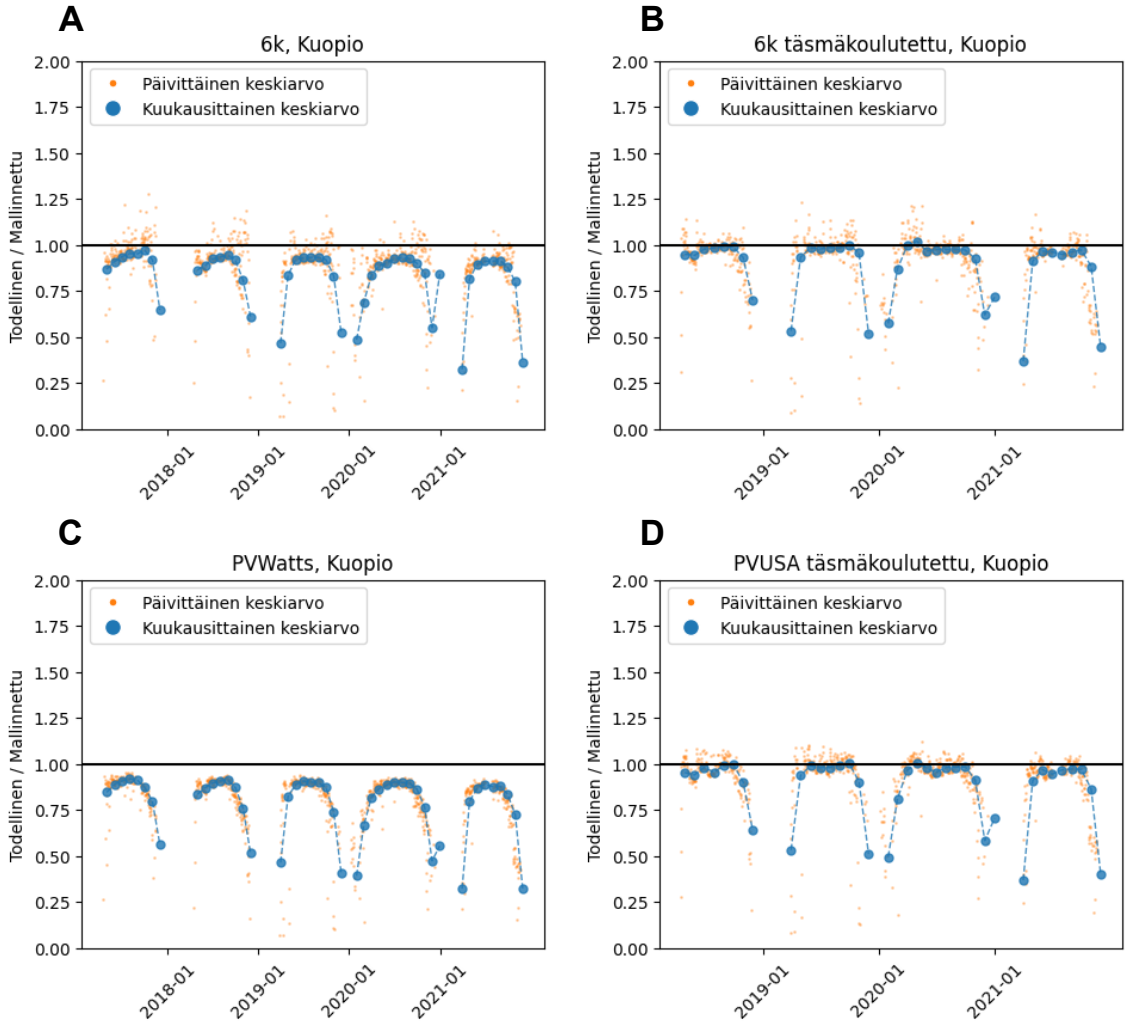
jolloin tehoprofiilissa esiintyy yllättäviä muutoksia ja sahalaitaisuutta. 6k-mallien kuvissa 5.3A ja 5.3B on esitetty tehoprofiilit yleiselle ja täsmäkoulutetulle 6k-mallille. Kuvista huomataan, että yleinen 6k-malli mallintaa tehon liian suureksi aurinkoisella päivällä. Käyrän alussa ja lopussa tapahtuu kuitenkin äkillisiä muutoksia tasaiseen käyrään verrattuna. Tämä voi johtua esimerkiksi paikallisesta pyranometrin peittävästä varjosta, joka ei kuitenkaan peitä kaikkia paneeleita. Paikallisesta pilvisyydestä voisi kertoa myös todellisen tehon pieni sahalaitaisuus käyrän loppupuolella, mikä voisi johtua siitä, että osa paneeleista on varjon peitossa, jolloin tehon tuotanto ei ole täysin tasaista. Täsmäkoulutetun 6k-mallin profiili on lähempänä todellista profiilia etenkin käyrän huipun kohdalla, mutta käyrän alku- ja loppupäätä kohden se alkaa yhtymään yleisen mallin ennustukseen. Pilvisen päivän kohdalla yleinen 6k-malli seuraa hyvin profiilin muotoa, mutta yliarvioi tehon tuotantoa monessa kohdassa. Yleisen 6k-mallin profiili noudattaa todellista profiilia erityisen hyvin klo 10 ja klo 12 välillä. Yleisen PVWatts-mallin mallintamat tehoprofiilit (kuvat 5.3C ja 5.3D) muistuttavat yleisen 6k-mallin profiileita. Täsmäkoulutetun PVUSA-mallin mallintamat tehoprofiilit (kuvat 5.3E ja 5.3F) muistuttavat puolestaan täsmäkoulutetun 6k-mallin profiileita. Tulokset ovat odotettuja, sillä taulukon 5.1 virhetermien ja pistekuvaajien 5.2 mukaan yleiset 6k- ja PVWatts-mallit vastaavat tarkkuudeltaan toisiaan, kun taas täsmäkoulutetut 6k- ja PVUSA-mallit vastaavat tarkkuudeltaan toisiaan.

Kuvassa 5.4 oransseilla pisteillä on esitetty mallinnettujen tehojen päivittäiset keskiarvot jaettuna todellisten tehojen päivittäisillä keskiarvoilla. Sinisillä pisteillä on esitetty puolestaan kuukausittaisten keskiarvojen suhde. Mikäli malli sopisi aineistoon täydellisesti, tulisi pisteiden olla y-akselin arvossa 1. Malleja tarkastellaan vain Kuopion aineistossa. Kuvista näkyy erityisesti kausittainen vaihtelu. Syksyllä pisteet alkavat painua alaspäin, mikä tarkoittaa sitä, että mallit mallintavat tehon tällöin liian suureksi. Talvisin aineistosta puuttuu havaintoja, mikä nähtiin myös puuttuvien arvojen kuvaajasta 4.1. Keväisin pisteet palaavat taas lähemmäs arvoa 1 ja kesäisin pisteet pysyttelevät melko lähellä arvoa 1. Tämä kausittainen vaihtelu johtuu siitä, että mallit mallintavat tehon arvon lähelle todellista tehoa kesäisin, kun säteilyä on enemmän, mutta talvisin esimerkiksi vähäisen säteilyn ja lumen takia mallit mallintavat tehon liian suureksi ja ovat näin ollen epäluotettavia talvisin. Kuvat vahvistavat aiempaa havaintoa siitä, että täsmäkouluttaminen parantaa malleja. Tulokset ovat niiltä osin yhteneväisiä Karttusen ym. tutkimuksen [6] kanssa, että mallien ennusteissa on kausittaista vaihtelua kesän ja talven välillä. Toisaalta Karttusen ym. tutkimuksessa PVUSA-malli näyttäisi aiheuttavan merkittävästi enemmän hajontaa kuin 6k-malli. Täsmäkoulutettujen 6k- ja PVUSA-mallien kuvaajat 5.4B ja 5.4D ovat melko lähellä toisiaan, mikä vahvistaa tämän työn toista aiempaa havaintoa, että täsmäkoulutetun 6k-mallin sekä täsmäkoulutetun PUVSA-mallin tarkkuus on lähellä toisiaan ainakin käytetyn aineiston osalta.

Kuvassa 5.5 on esitetty mallinnettu energian tuotanto ja todellinen energian tuotanto kullekin kuukaudelle Kuopion järjestelmässä. Energiatuotanto on laskettu kaavalla 2.3, jos-

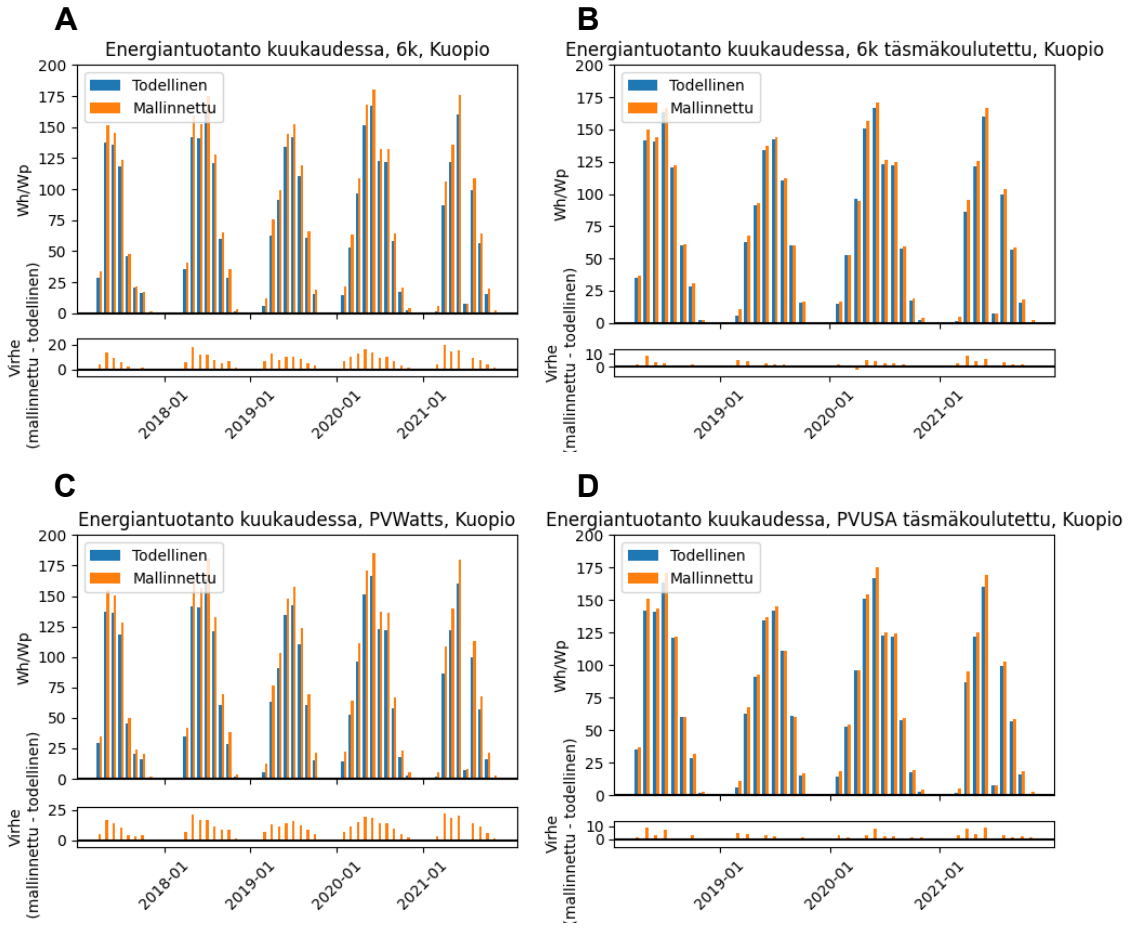


Kuva 5.3. *Fysikaalisten mallien tehoprofiilit Kuopion aineistossa. Täsmäkoulutetun 6k-mallin ja PVUSA-mallin kouluttamiseen käytetään Kuopion järjestelmän aineiston ensimmäiset 365 päivää.*



Kuva 5.4. *Fysikaalisten mallien todellisten ja mallinnettujen tehojen suhde päivä- ja kuukausitasolla Kuopion järjestelmässä. Täsmäkoulutetun 6k-mallin ja PVUSA-mallin kouluttamiseen käytetään Kuopion järjestelmän aineiston ensimmäiset 365 päivää.*

sa k käy läpi kaikki kuukauden mitatut havainnot, P_k on jokaista havaintoa k vastaava tehon arvo ja t_k on jokaisen mittausjakson pituus, eli tässä tapauksessa 1/60, sillä havainnot ovat minuutin resoluutiossa. Saatu kuukauden energia jaetaan vielä järjestelmän nimellisteholla, jolloin yksiköksi saadaan Wh/Wp. Kunkin kuvan alla on esitetty myös kuukauden virhe. Positiivinen virheen arvo tarkoittaa, että mallinnettu arvo on suurempi kuin todellinen arvo ja negatiivinen virheen arvo tarkoittaa, että mallinnettu arvo on pienempi kuin todellinen arvo. Kuvista huomataan, että energian tuotanto on huomattavasti korkeampaa kesäisin ja laskee talvisin. Yleisten 6k- ja PVWatts-mallien kuvaajista 5.5A ja 5.5C nähdään, että nämä mallit mallintavat energian tuotannon systemaattisesti liian suureksi. Tämä on seurausta siitä, että nämä mallit mallintavat tehon lähes systemaattisesti liian suureksi. Täsmäkoulutuksen myötä energian tuotannon mallinnuksen virhe pienenee, mutta virhe näyttäisi silti olevan pääsääntöisesti positiivista. Tämä tarkoittaa sitä, että vaikka täsmäkoulutus vähentää mallien taipumusta mallintaa energian tuotanto liian suureksi, näyttäisi malleille silti jäävän pieni taipumus mallintaa energian tuotanto liian



Kuva 5.5. *Fysikaalisten mallien todelliset ja mallinnetut kuukausittaiset energiatuotannot Kuopion järjestelmässä. Täsmäkoulutetun 6k-mallin ja PVUSA-mallin kouluttamiseen käytetään Kuopion järjestelmän aineiston ensimmäiset 365 päivää.*

suureksi. Etenkin yleisten 6k- ja PVWatts-mallien kuvaajista 5.5A ja 5.5C huomataan, että virhe pienenee talvisin ja kasvaa kesällä. Tämä voisi näyttää olevan ristiriidassa kuvan 5.4 kanssa, jossa suhteellinen virhe kasvaa talvisin. Ristiriidalta näyttävä havainto johtuu kuitenkin siitä, että talvisin hetkittäinen teho ja energian tuotanto ovat pienempiä kuin kesäisin, joten vaikka absoluuttinen virhe olisikin pieni, voi se olla iso suhteessa tuotettuun energiaan. Esimerkiksi 100 W absoluuttinen ero on pieni, kun sitä verrataan kesäpäivään, jolloin hetkittäinen teho on 10000 W, mutta suuri, kun sitä verrataan talvipäivään, jolloin hetkittäinen teho on 200 W.

Taulukossa 5.2 on esitetty vuosittaiset energian tuotannon mallinnusten virheet eri fysikaalisille malleille. Aiemmista kuvaajista poiketen taulukossa 5.2 on esitetty virheet kaikille järjestelmille. Taulukkoon on otettu vuodet Kuopion aineiston perusteella, vaikka esimerkiksi Helsingin järjstelmässä aineistoa on vuodesta 2015 vuoteen 2024. Positiivinen arvo tarkoittaa, että energian tuotanto on mallinnettu todellisuutta suuremmaksi ja negatiivinen arvo tarkoitaa, että energian tuotanto on mallinnettu todellisuutta pienemmäksi. Koska 6k-mallin ja PVUSA-mallin täsmäkoulutuksiin käytetään aineiston 365 ensimmäistä päivää, ei aineistojen ensimmäisen vuoden virheitä lasketa täsmäkoulutetuille malleil-

le. Koska aineisto ei välttämättä ala vuoden alusta, täsmäkoulutukseen voidaan käyttää myös seuraavan vuoden aineistoa. Tämä tarkoittaa sitä, että toisen vuoden aineisto saattaa poiketa yleisten ja täsmäkoulutettujen mallien välillä, jolloin toisen vuoden virheet eivät ole välttämättä suoraan vertailukelpoisia. Vuosittaiset virheet vaihtelevat suuresti järjestelmästä riippuen. Esimerkiksi Sodankylän aineistossa yleiset mallit mallintavat vuosittaisen energian tuotannon noin 20–30 % liian suureksi. Sodankylän järjestelmien suuret virheet voivat johtua esimerkiksi muita järjestelmiä pohjoisemmasta sijainnista, jolloin järjestelmät ovat alttiimpia lumelle, mutta tarkkaa syytä on vaikea sanoa ilman yksityiskohtaisempaa tarkastelua. Mallien täsmäkoulutus pienentää virhetermien arvoja jokaisessa järjestelmässä. Vuosittainen energian tuotannon mallinnuksen virhe ei kuitenkaan välttämättä kuvaa parhaiten täsmäkoulutuksen vaikutusta, sillä yleiset mallit näyttävät systemaattisesti mallintavan tehon liian suureksi, jolloin virhe kumuloituu vuosittaiseen energian tuotantoon. Täsmäkoulutuksen myötä mallien systemaattinen taipumus mallintaa tehoa liian suureksi vähenee ja osa tehon arvoista mallintuukin liian pieneksi, jolloin osa virheistä kumoutuu. Tällöin vuosittaiset energian tuotannon virheetkin pienenevät.

5.2 Koneoppimismallien tulokset

Koneoppimismallit on muodostettu kappaleessa 4.3 esitetyllä tavalla. Taulukossa 5.3 on esitetty koneoppimismallien R^2-arvot ja virhetermit eri järjestelmille. Kuopion järjestelmä on malleille aivan uusi järjestelmä, jonka avulla tarkastellaan mallien toimivuutta uudessa järjestelmässä. Vaikka Helsingin, Sodankylän ja Turun järjestelmiä käytettiin mallien kouluttamiseen, on koulutusaineistosta jätetty vuosi 2019 pois. Tällöin mallien toimintaa voidaan tarkastella myös koulutukseen käytetyissä järjestelmissä, mutta kuitenkin sellaisella aineistolla, jota ei ole käytetty koulutukseen. Taulukkoon 5.3 on merkitty myös fysikaalinen 6k-malli vertailun helpottamiseksi. Kun tarkastellaan yleisiä malleja Kuopion järjestelmässä, HGBR malli tuottaa suurimman R^2-termin arvon ja pienimmät virhetermit. Myös yleinen MLP-malli näyttäisi virhetermien perusteella toimivan hieman paremmin, kuin yleinen 6k-malli. Kun tarkastellaan täsmäkoulutusta Kuopion järjestelmässä, MLP malli tuottaa suurimman R^2-termin arvon ja pienemmät virhetermit. HGBR-malli ei puolestaan parane täsmäkoulutuksessa virhetermien perusteella. HGBR-malliin sovitetaan vain yksi uusi heikko oppija täsmäkoulutuksessa aikaisen lopetuksen takia. On kuitenkin mahdollista, että täsmäkoulutuksen tulos paranisi, jos aikaista lopetusta ei käytettäisi ja annettaisiin täsmäkoulutuksen luoda malliin useampia heikkoja oppijoita. Taulukon tulosten myötä koneoppimismallit näyttäisivät toimivan hieman fysikaalisia malleja paremmin sekä yleisinä että täsmäkoulutettuina malleina. Ero on kuitenkin aika pieni ja toisella järjestelmällä tulokset voisivat olla erilaisia.

Koneoppimismallit antavat Helsingin järjestelmässä vuodelle 2019 hyvin pieniä virhearvoja ja korkean R^2-termin, mikä on odotettu tulos sillä Helsingin järjestelmän aineistoa

Taulukko 5.2. *Fysikaalisten mallien vuosittaiset energian tuotannon virheet (prosentteina). Taulukkoon on valittu vuodet, jotka löytyvät Kuopion aineistosta. Koska 6k-mallin ja PVUSA-mallin täsmäkouluttamiseen käytetään aineiston ensimmäiset 365 päivää, puuttuu tämän takia kyseisistä malleista aikaisin vuosi. Koska Helsingin aineisto alkaa jo vuodesta 2015, ei koulutus aiheuta taulukkoon Helsingin aineistoon puuttuvia arvoja. Positiivinen arvo tarkoittaa, että energian tuotanto on mallinnettu todellista tuotantoa suuremmaksi ja negatiivinen arvo tarkoittaa, että energian tuotanto on mallinnettu todellista tuotantoa pienemmäksi.*

Järjestelmä	Malli	2017	2018	2019	2020	2021
Kuopio (IL)	6k (yleinen)	$7,7$	$9,8$	10	11	14
	6k (täsmäkoulutettu)	-	$3,2$	$3,1$	$2,7$	$5,8$
	PVWatts (yleinen)	12	13	14	15	18
	PVUSA (täsmäkoulutettu)	-	$3,9$	$3,0$	$3,3$	$6,2$
Helsinki (IL)	6k (yleinen)	11	11	14	15	16
	6k (täsmäkoulutettu)	$-1,7$	$1,5$	$0,3$	$1,7$	$1,0$
	PVWatts (yleinen)	15	15	17	18	19
	PVUSA (täsmäkoulutettu)	$-3,3$	$5,2$	$1,6$	$2,5$	$4,9$
Sodankylä 20 astetta (IL)	6k (yleinen)	-	24	27	29	32
	6k (täsmäkoulutettu)	-	-	$5,5$	$7,3$	$9,0$
	PVWatts (yleinen)	-	26	31	32	35
	PVUSA (täsmäkoulutettu)	-	-	$1,7$	$3,4$	$6,4$
Sodankylä 90 astetta (IL)	6k (yleinen)	-	19	25	24	27
	6k (täsmäkoulutettu)	-	-	$5,3$	$7,1$	$9,1$
	PVWatts (yleinen)	-	23	30	27	31
	PVUSA (täsmäkoulutettu)	-	-	$5,3$	$4,8$	$7,9$
Turku (Turku AMK)	6k (yleinen)	-	13	14	20	-
	6k (täsmäkoulutettu)	-	-	$1,1$	$6,6$	-
	PVWatts (yleinen)	-	16	17	23	-
	PVUSA (täsmäkoulutettu)	-	-	$0,1$	$5,4$	-

on käytetty mallien kouluttamiseen. Vaikka Sodankylän 20 asteen järjestelmää ja Turun järestelmää on käytetty myös mallien kouluttamiseen, niiden virhetermit ovat Helsingin ja Sodankylän 90 asteen järjestelmiä selvästi suurempia vuonna 2019. Poikkeavan suuret virheet voivat johtua näiden järjestelmien aineistojen kohinasta, joka näkyy kuvissa 5.1C ja 5.1E. Koska Helsingin järjestelmän aineisto kattaa noin 9 vuotta, Sodankylän järjestelmät 6 vuotta ja Turun aineisto vain 3 vuotta, on mallien kouluttamiseen käytetty eniten Helsingin järjestelmän aineistoa. Tämä todennäköisesti johtaa siihen, että mallit toimivat erityisen hyvin Helsingin aineistossa, sillä järjestelmä on malleille koulutuksen myötä tutuin.

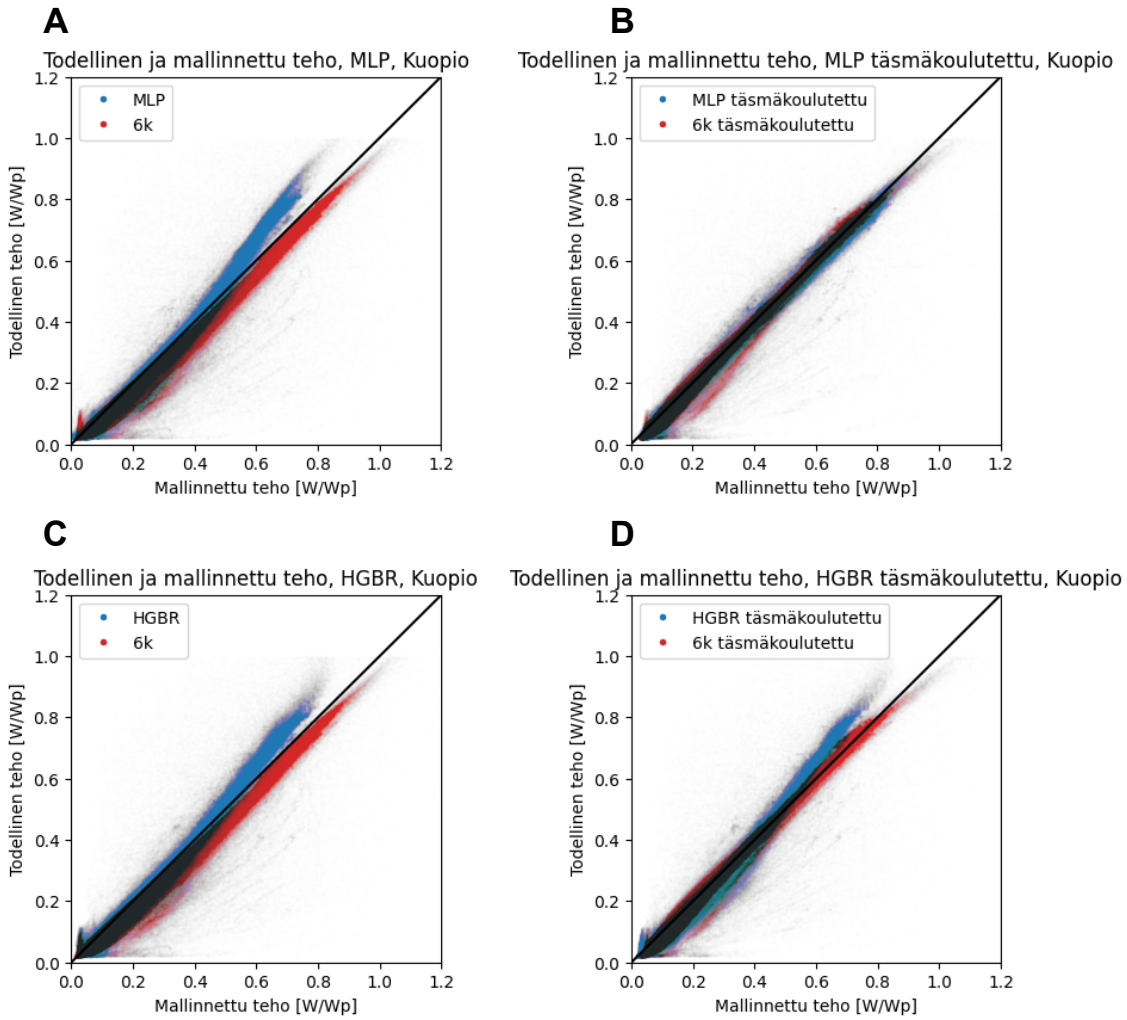
Kuvassa 5.6 on esitetty pistekuvaajat, joista näkyy miten hyvin koneoppimismalleilla mallinnetut tehot vastaavat todellisia tehon arvoja. Kuviin on lisätty 6k-mallin pistekuvaa-

Taulukko 5.3. *Koneoppimismallien virhetermit. Yleisten MLP- ja HGBR-mallien virhetermit on laskettu suoraan Kuopion aineistoon. Täsmäkoulutetun MLP-mallin ja HGBR-mallin kouluttamiseen käytetään Kuopion järjestelmän aineiston ensimmäiset 365 päivää, ja virhetermit lasketaan jäljelle jääneeseen aineistoon. Vaikka MLP- ja HGBR-mallit koulutetaan Helsingin, Sodankylän ja Turun järjestelmien aineistoilla, vuotta 2019 ei olla käytetty kouluttamiseen. Taulukossa on esitetty myös fysikaalinen 6k-malli vertailun vuoksi.*

Järjestelmä	Malli	R^2	MAE	RMSE
Kuopio (IL) 2017–2021	MLP (yleinen)	$0,94$	$0,04$	$0,06$
	MLP (täsmäkoulutettu)	$0,98$	$0,02$	$0,04$
	HGBR (yleinen)	$0,95$	$0,03$	$0,05$
	HGBR (täsmäkoulutettu)	$0,95$	$0,03$	$0,06$
	6k (yleinen)	$0,93$	$0,04$	$0,06$
	6k (täsmäkoulutettu)	$0,96$	$0,03$	$0,05$
Helsinki (IL) 2019	MLP	$0,99$	$0,01$	$0,02$
	HGBR	$0,99$	$0,01$	$0,02$
Sodankylä 20 astetta (IL) 2019	MLP	$0,89$	$0,04$	$0,06$
	HGBR	$0,89$	$0,03$	$0,06$
Sodankylä 90 astetta (IL) 2019	MLP	$0,98$	$0,01$	$0,03$
	HGBR	$0,98$	$0,01$	$0,03$
Turku (Turku AMK) 2019	MLP	$0,96$	$0,03$	$0,05$
	HGBR	$0,96$	$0,03$	$0,05$

jat helpottamaan vertailua. Yleisten MLP- ja HGBR-mallien kuvaajat 5.6A ja 5.6C ovat samankaltaisia. Kumpikin malli näyttäisi mallintavan alhaiset tehon arvot liian suureksi, mutta suuremmat tehon arvot liian pieneksi. Kun pisteiden muotoa verrataan yleiseen 6k-malliin, huomataan hyvin miten pisteet alkavat eroamaan toisistaan tehon kasvaessa. Täsmäkoulutuksen myötä erityisesti MLP-mallin tuottamat pisteet asettuvat paremmin suoralle kuvaajan 5.6B perusteella. Täsmäkoulutetun HGBR-mallin tuottamat pisteet näyttäisivät vastaavan yleisen HGBR-mallin pisteitä kuvaajan 5.6D perusteella.

Kuvassa 5.7 on esitetty MLP- ja HGBR-mallien tehoprofiilit samoilta päiviltä kuin kuvassa 5.3. Kuvaajiin on lisätty myös yleisen ja täsmäkoulutetun 6k-mallin tehoprofiilit helpottamaan fysikaalisten ja koneoppimismallien vertailua. Kuvaajista 5.7A ja 5.7C huomataan, että aurinkoisena päivänä yleiset MLP- ja HGBR-mallit, sekä täsmäkoulutettu HGBR-malli, mallintavat tehon liian pieneksi päivällä, kun teho on korkeimmillaan. Myös pilvisen päivän kuvaajissa 5.7B ja 5.7D yleiset MLP- ja HGBR-mallit, sekä täsmäkoulutettu HGBR-malli, mallintavat tehon liian pieneksi korkeimmilla tehon arvoilla. Esimerkiksi



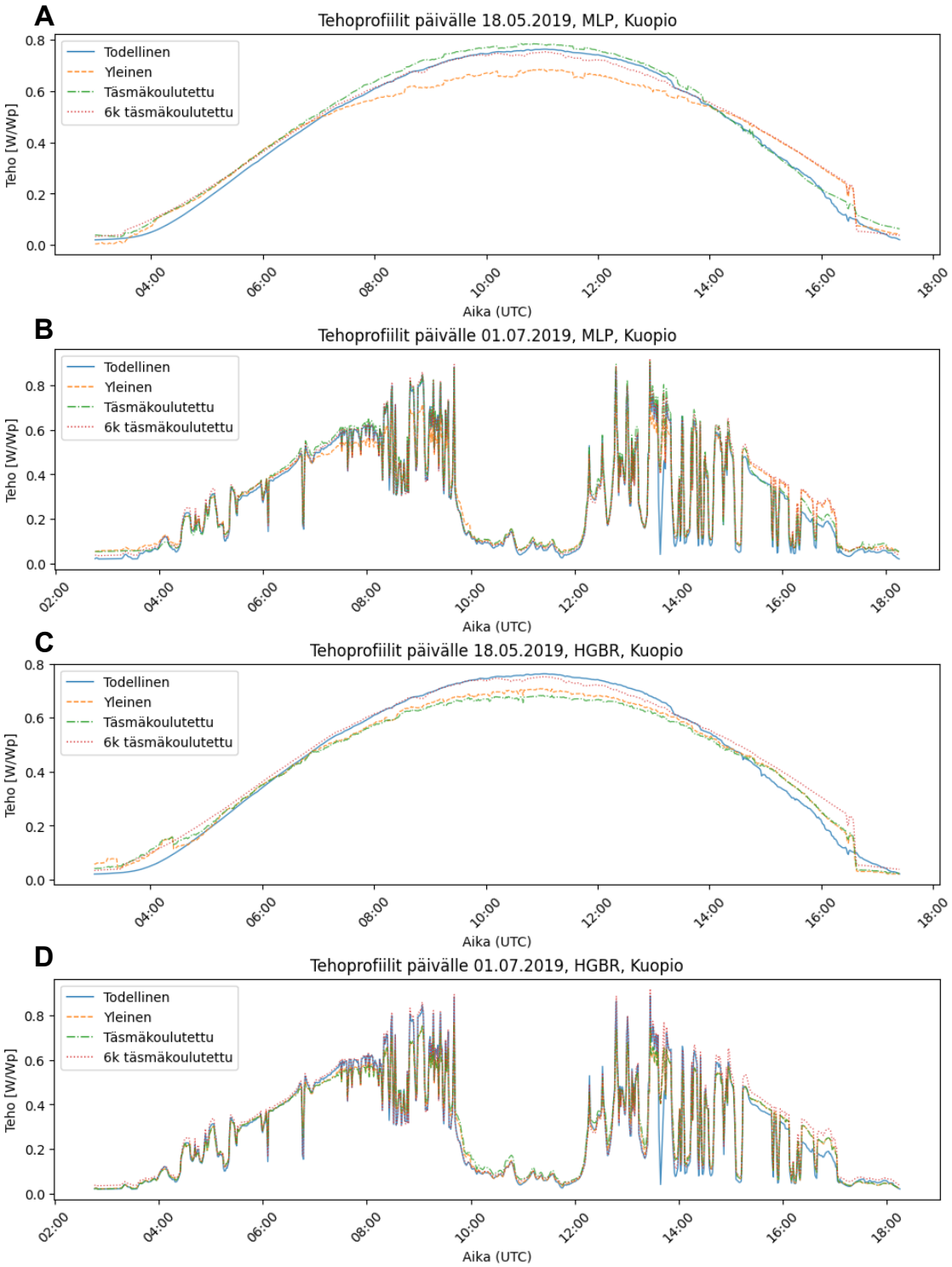
Kuva 5.6. *Koneoppimismallien todellisten ja mallinnettujen tehojen pistekuvaajat Kuopion järjestelmässä. Täsmäkoulutetun MLP-mallin ja HGBR-mallin kouluttamiseen käytetään Kuopion järjestelmän aineiston ensimmäiset 365 päivää. Kuvaajiin on lisätty fysikaalinen 6k-malli helpottamaan vertailua.*

kuvaajasta 5.7B nähdään, että yleisen MLP-mallin tehoprofiili on alhaisempi kuin todellinen tehoprofiili kello 9 aikoihin. Koneoppimismallien taipumus mallintaa suurimmat tehon arvot liian pieneksi näkyi myös pistekuvaajissa 5.6, joissa yleisten MLP- ja HGBR-mallien pisteet nousivat diagonaalin yläpuolelle suuremmilla tehoilla. MLP-mallin täsmäkoulutus parantaa sovitusta huomattavasti. Kuvaajasta 5.7A nähdään, että aurinkoisena päivänä täsmäkoulutettu MLP-malli seuraa illan sahalaitaista tehoprofiilia paremmin kuin täsmäkoulutettu 6k-malli. Kuvaajasta 5.7C nähdään, että myös täsmäkoulutetun HGBR-mallin tehoprofiili on aurinkoisella päivällä myös hieman todellista tehoprofiilia leveämpi ja matalampi. Yleisessä ja täsmäkoulutetussa HGBR-mallissa on havaittavissa todellisesta tehoprofiilista poikkeavaa rakennetta klo 3 ja klo 5 välillä, mitä muissa malleissa ei näy näin paljon. Kuvaajasta 5.7C nähdään, että HGBR-mallin täsmäkoulutus näyttää jopa hieman huonontavan mallinnettua tehoprofiilia, sillä käyrä näyttäisi madaltuvan entisestään ja näin ollen jäävän kauemmas todellisesta tehoprofiilista. Kuvaajasta 5.7D nähdään, et-

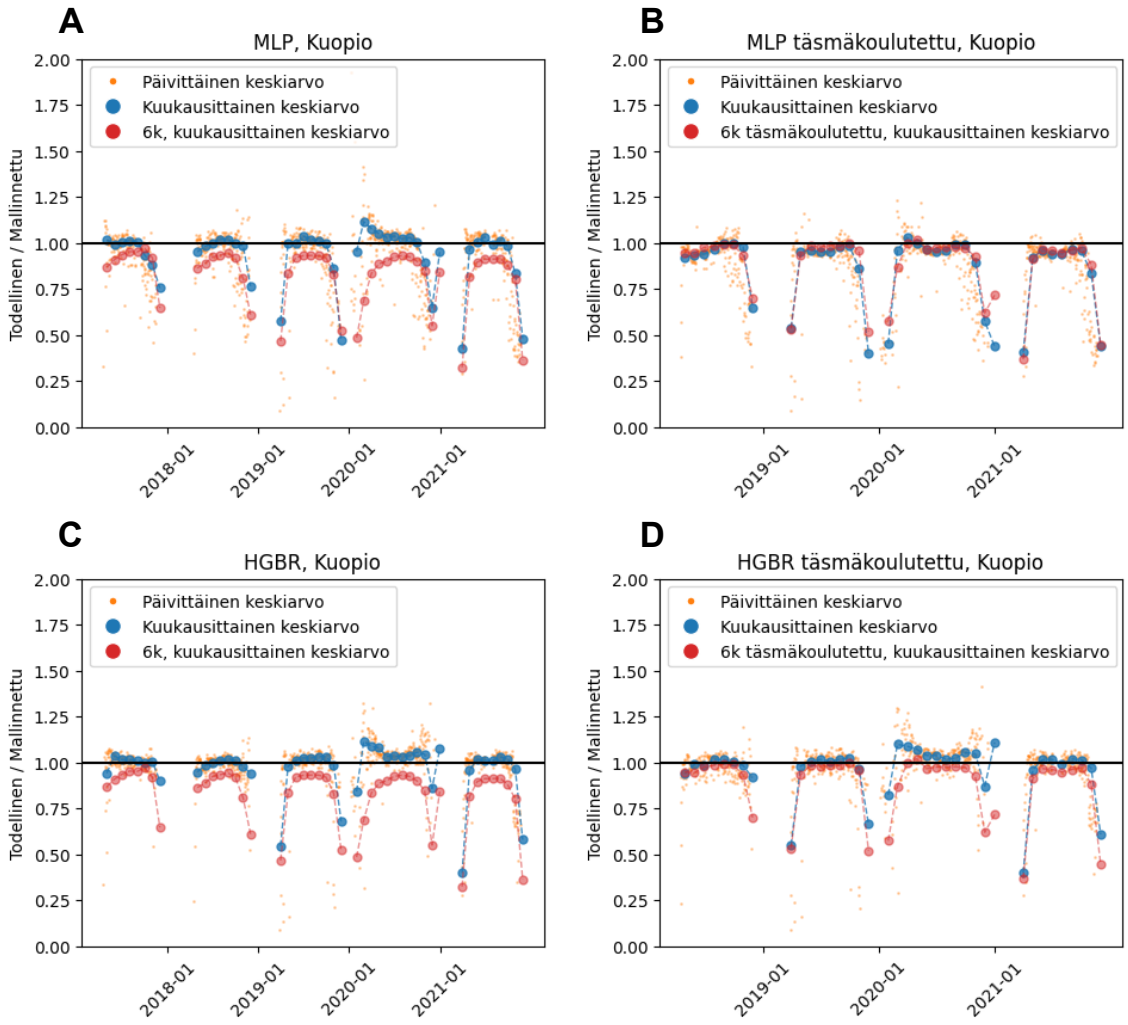
tä yleinen ja täsmäkoulutettu HGBR-malli näyttäisi kuitenkin noudattavan pilvisen päivän tehoprofiilia aika hyvin. Täsmäkoulutus näyttäisi hieman huonontavan HGBR-mallin tehoprofiilia myös pilvisellä säällä, sillä täsmäkoulutetun mallin tehoprofiili poikkeaa etenkin aikavälillä klo 10–12 toteutuneesta tehoprofiilista enemmän kuin yleisen HGBR-mallin tehoprofiili.

Kuvassa 5.8 on esitetty koneoppimismalleille samat kuvaajat, kuin fysikaalisten mallien kuvaajat kuvassa 5.4. Kuvaajiin on lisätty 6k-mallin tuottamat keskiarvojen suhteet punaisella helpottamaan fysikaalisten ja koneoppimismallien vertailua. Yleisten MLP- ja HGBR-mallien kuvaajissa 5.8A ja 5.8C huomataan selkeä ero fysikaaliseen 6k-malliin verrattuna. Molempien yleisten koneoppimismallien kuukausittaist pisteet ovat keskimäärin lähempänä arvoa 1, kuin yleisen 6k-mallin pisteet, mikä viittaa tarkempiin ennusteisiin. MLP-mallin täsmäkoulutus näyttäisi vähentävän hieman pisteiden hajontaa päivittäisissä ja kuukausittaisissa keskiarvojen suhteissa kuvaajan 5.8B perusteella. Täsmäkoulutetun MLP- ja 6k-mallin kuukausittaiset pisteet näyttävät olevan aika lähellä toisiaan kuvaajan 5.8B perusteella. Kun täsmäkoulutetun MLP- ja 6k-mallien kuvaajia 5.8B ja 5.4B verrataan toisiinsa, näyttäisi siltä, että täsmäkoulutetun 6k-mallin päivittäisten keskiarvojen suhteissa on enemmän hajontaa arvon 1 ympärillä, kuin täsmäkoulutetun MLP-mallilla. Ero on kuitenkin aika pientä ja näyttäisi siltä, että täsmäkoulutetut MLP- ja 6k-mallit antavat samankaltaisia tuloksia. Täsmäkoulutetun HGBR-mallin päivittäisen ja kuukausittaisten keskiarvojen suhteet eivät näytä poikkeavan kovinkaan paljon yleisen HGBR-mallin pisteistä, kun kuvaajia 5.8C ja 5.8D vertaillaan.

Kuvassa 5.9 on esitetty koneoppimismalleilla mallinnettu ja oikea energiatuotanto kullekin kuukaudelle Kuopion järjestelmässä. Kuvan tulkinta on samanlainen kuin vastaavassa fysikaalisten mallien kuvassa 5.5. Yleisten MLP- ja HGBR-mallien kuvaajissa 5.9A ja 5.9C mallinnetut ja todelliset kuukausittaiset energian tuotannot vastaavat melko hyvin toisiaan. Vuosi 2020 näyttäisi tuottavan kummallekin yleiselle mallille eniten virhettä. Kumpikin yleinen malli mallintaa kuukausittaiset energian tuotannot systemaattisesti liian pieneksi vuonna 2020. Muina vuosina yleisillä malleilla näyttäisi olevan lievä taipumus mallintaa kesäkuukausien energian tuotannon liian pieneksi, mutta muina vuodenaikoina liian suureksi. Tulos on linjassa kuvaajien 5.6A ja 5.6C kanssa, joissa on esitetty yleisten koneoppimismallien mallinnettujen ja todellisten tehon arvojen pistekuvaaja. Näissä kuvaajissa näkyi yleisten koneoppimismallien taipumus mallintaa pienet tehon arvot liian suuriksi ja suuret tehon arvot puolestaan liian pieniksi. Nämä tehon virheet näyttäisivät siis kumuloituvan kuukausitasolla. Täsmäkoulutettu MLP-malli näyttäisi kuvaajan 5.9B mukaan mallintavan kuukausittaisen energian tuotannon melko systemaattisesti liian suureksi. Tulos on yllättävä, sillä aikaisempien kuvaajien ja virhetermien mukaan täsmäkoulutus tekisi MLP-mallista tarkemman. Tämä voi johtua esimerkiksi siitä, että täsmäkoulutuksen myötä MLP-mallin taipumus mallintaa pienet tehon arvot liian suureksi ja suuret tehon arvot liian pieneksi vähenee, jolloin virheet eivät kumoa toisiaan enää niin paljon kuukausita-



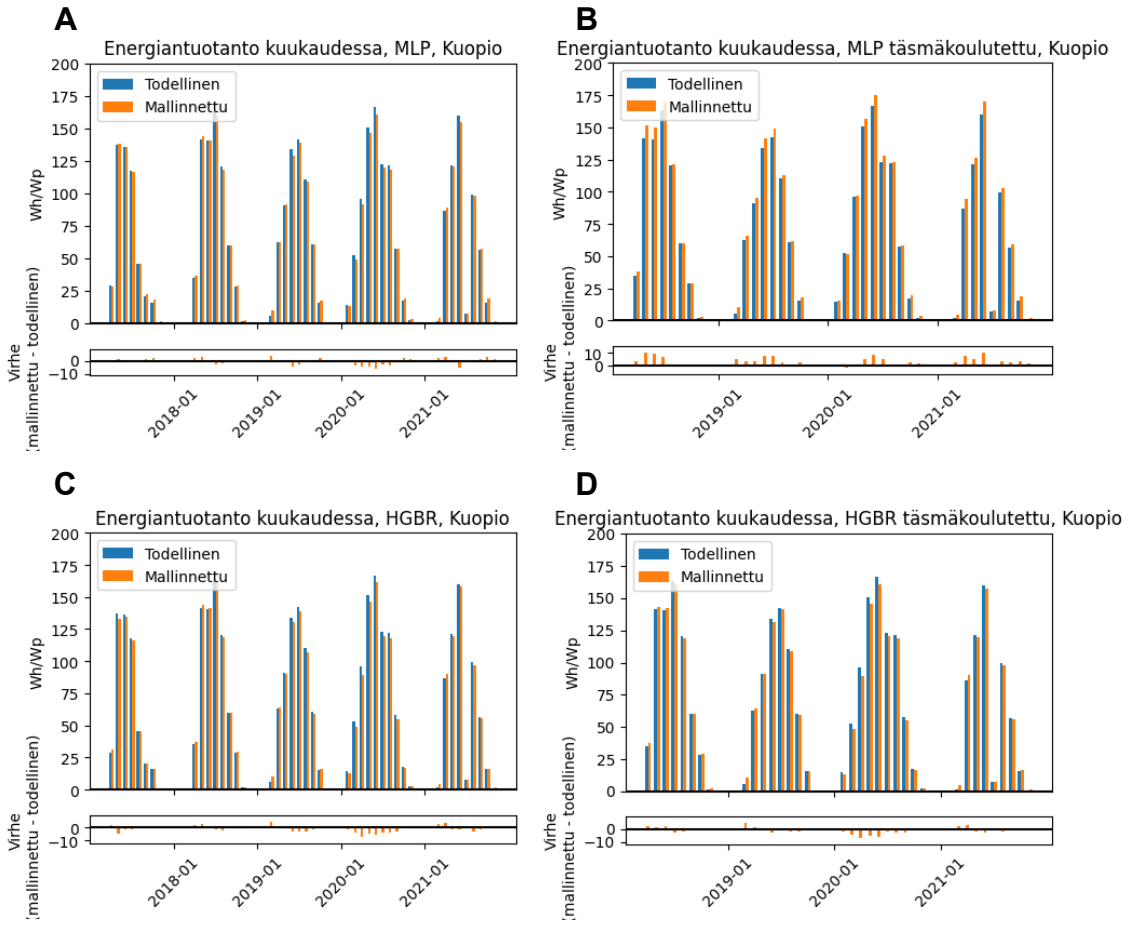
Kuva 5.7. *Koneoppimismallien tehoprofiilit Kuopion aineistossa. Täsmäkoulutetun MLP-mallin ja HGBR-mallin kouluttamiseen käytetään Kuopion järjestelmän aineiston ensimmäiset 365 päivää. Kuvaajiin on lisätty myös 6k-mallin tehoprofiili helpottamaan vertailua.*



Kuva 5.8. *Koneoppimismallien todellisten ja mallinnettujen tehojen suhde päivä- ja kuukausitasolla Kuopion järjestelmässä. Täsmäkoulutetun MLP-mallin ja HGBR-mallin kouluttamiseen käytetään Kuopion järjestelmän aineiston ensimmäiset 365 päivää. Kuvaajiin on lisätty fysikaalinen 6k-malli helpottamaan vertailua.*

solla. Täsmäkoulutus ei näytä vaikuttavan HGBR-mallin virheisiin suuresti kuvaajan 5.9D perusteella.

Taulukossa 5.4 on esitetty vuosittaiset energian tuotannon virheet eri koneoppimismalleille kaikissa järjestelmissä. 6k-mallin vuosittaiset virheet on lisätty taulukkoon helpottamaan vertailua. Koska Helsingin, Sodankylän ja Turun aineistoja vuotta 2019 lukuun ottamatta on käytetty koneoppimismallien kouluttamiseen, jätetään nämä vuodet pois taulukosta. Kuten fysikaalisten mallien tapauksessa taulukossa 5.2, koneoppimismallienkin täsmäkoulutukseen käytetään aineiston 365 ensimmäistä päivää. Täsmäkoulutetuille malleille ei lasketa virheitä siis ensimmäiseltä vuodelta. Lisäksi toisen vuoden virheet eivät välttämättä ole tässäkään vertailukelpoisia, sillä osa toisesta vuodesta on mahdollisesti myös käytetty täsmäkoulutukseen. Positiivinen virhe tarkoittaa, että mallinnettu energian tuotanto on todellista energian tuotantoa suurempaa. Negatiivinen virhe tar-



Kuva 5.9. *Koneoppimismallien todelliset ja mallinnetut kuukausittaiset energiatuotannot Kuopion järjestelmässä. Täsmäkoulutetun MLP-mallin ja HGBR-mallin kouluttamiseen käytetään Kuopion järjestelmän aineiston ensimmäiset 365 päivää.*

koittaa taas, että mallinnettu energian tuotanto on todellista energian tuotantoa pienempää. Yleisten koneoppimismallien vuosittaiset virheet ovat suuruudeltaan huomattavasti pienempiä, kuin yleisen 6k-mallin. MLP-mallin täsmäkoulutus lisää kuitenkin vuosittaisia virheitä. Koska täsmäkoulutetun MLP-mallin kuukausittaisten energian tuotannon virheiden nähtiin kasvavan kuvaajassa 5.9B, vuosittaisten energian tuotannon virheiden kasvu ei ole yllättävä tulos. HGBR-mallin täsmäkoulutus ei näytä vähentävän vuosittaisia energian tuotannon mallinnuksen virheitä. Kuvista 5.8 ja 5.9 huomattiin, että vuosi 2020 näytti tuottavan eniten virheitä etenkin yleisillä koneoppimismalleilla. Virheet näyttäisivät kumuloituvan myös vuosittaisiin energian tuotantoihin, sillä taulukossa 5.4 Kuopion suurimmat virheet yleisille koneoppimismalleille tulee vuonna 2020. Helsingin, Sodankylän ja Turun virheet ovat kummallakin mallilla alle 2 %. Pienet virheet näillä aineistoilla on odotettu tulos, sillä vaikka kyseisiä aineistoja ei olla käytetty mallien koulutukseen, on aineistoa näistä järjestelmistä käytetty mallien koulutukseen. Tämä johtaa siihen, että mallit ovat saattaneet oppia näiden järjestelmien ominaispiirteitä.

Taulukko 5.4. *Koneoppimismallien vuosittaiset energian tuotannon virheet (prosentteina). Taulukkoon on valittu vuodet, jotka löytyvät Kuopion aineistosta. Koska MLP-mallin ja HGBR-mallin täsmäkouluttamiseen Kuopion järjestelmään käytetään aineiston ensimmäiset 365 päivää, puuttuu tämän takia Kuopion järjestelmästä täsmäkoulutetuista malleista aikaisin vuosi. Mallien kouluttamiseen on käytetty Helsingin, Sodankylän ja Turun aineistoja lukuun ottamatta vuotta 2019, joten näistä ei lasketa virheitä. Taulukkoon on lisätty myös 6k-malli helpottamaan koneoppimismallien vertailua fysikaalisiin malleihin. Positiivinen arvo tarkoittaa, että energian tuotanto on mallinnettu todellista tuotantoa suuremmaksi ja negatiivinen arvo tarkoittaa, että energian tuotanto on mallinnettu todellista tuotantoa pienemmäksi.*

Järjestelmä	Malli	2017	2018	2019	2020	2021
Kuopio (IL)	MLP (yleinen)	$0,5$	$0,2$	$-0,1$	$-2,8$	$0,8$
	MLP (täsmäkoulutettu)	-	$4,5$	$5,2$	$3,0$	$6,6$
	HGBR (yleinen)	$-1,3$	$0,1$	$-1,0$	$-4,2$	$0,1$
	HGBR (täsmäkoulutettu)	-	$0,3$	$0,1$	$-3,9$	$0,2$
	6k (yleinen)	$7,7$	$9,8$	10	11	14
	6k (täsmäkoulutettu)	-	$3,2$	$3,1$	$2,7$	$5,8$
Helsinki (IL)	MLP	-	-	$0,5$	-	-
	HGBR	-	-	$-1,1$	-	-
	6k (täsmäkoulutettu)	$-1,7$	$1,5$	$0,3$	$1,7$	$1,0$
Sodankylä 20 astetta (IL)	MLP	-	-	$1,0$	-	-
	HGBR	-	-	$1,5$	-	-
	6k (täsmäkoulutettu)	-	-	$5,5$	$7,3$	$9,0$
Sodankylä 90 astetta (IL)	MLP	-	-	$-1,1$	-	-
	HGBR	-	-	$1,9$	-	-
	6k (täsmäkoulutettu)	-	-	$5,3$	$7,1$	$9,1$
Turku (Turku AMK)	MLP	-	-	$-1,5$	-	-
	HGBR	-	-	$-1,2$	-	-
	6k (täsmäkoulutettu)	-	-	$1,1$	$6,6$	-

5.3 Jatkokysymykset

Koska työssä käytettiin PV-aineistoa vain neljästä eri sijainnista ja viidestä eri järjestelmästä, aineisto kattaa vain pienen osan mahdollisista järjestelmien kokoonpanoista. Täten tässä työssä luotujen mallien tarkkuus saattaa olla huonompi PV-järjestelmillä, joiden rakenne poikkeaa tämän työn aineistosta paljon. Koneoppimismallien toimintaa testattiin vain yhdessä uudessa järjestelmässä, mikä ei vielä anna luotettavaa tulosta mallien toiminnasta. Malleja voisi siis testata useammilla PV-järjestelmillä – mahdollisuuksien mukaan myös muista Pohjoismaista.

Koska työn koneoppimismalleissa käytetään 20:tä syötemuuttujaa, voi mallien käyttöönotto olla hankalaa. Käyttöönoton helpottamiseksi olisi tärkeää tarkastella mitkä syötemuut-

tujat ovat tehon mallintamisessa merkitsevimpiä ja mitkä syötemuuttujat voitaisiin jättää pois malleista.

6. YHTEENVETO

Tässä työssä tutkittiin fysikaalisten tehomallien ja koneoppimispohjaisten tehomallien aiheuttamaa hetkittäisen tehon mallinnuksen virhettä Suomessa suureiden R^2, MAE ja RMSE avulla. Myös malleista aiheutuvaa energian tuotannon mallinnuksen virhettä tutkittiin kuukausi- ja vuositasolla vertaamalla mallinnettua energian tuotantoa toteutuneeseen energian tuotantoon. Käytettyjä fysikaalisia malleja olivat 6k, PVWatts ja PUVSA. Käytettyjä koneoppimismalleja olivat neuroverkkopohjainen monikerroksinen päättelin (*multilayer perceptron*, MLP) ja histogrammipohjainen gradienttiboostausmalli (*histogram gradient boosting regressor*, HGBR). Malleja tarkasteltiin sekä käyttämällä niitä suoraan PV-järjestelmiin yleisenä mallina ilman erillistä kouluttamista että täsmäkouluttamalla ne PV-järjestelmiin käyttämällä osaa aineistosta.

Ensimmäisenä tutkimuskysymyksenä oli tarkastella fysikaalisten mallien virheitä hetkittäisen tehon mallinnuksessa sekä energian tuotannon kuukausi- ja vuositason mallinnuksessa. Mallien hetkittäisen tehon virheet vaihtelivat eri PV-järjestelmien välillä. Erityisesti Sodankylän PV-järjestelmä, jossa paneelit on asennettu 20 asteen kulmaan, poikkesi muista järjestelmistä mallien suurten virhetermien perusteella. Yleisesti ottaen mallien virheet pienenivät ja R^2-arvo suureni, kun malleja täsmäkoulutettiin. Työssä tutkituista fysikaalisista malleista 6k-malli oli kuitenkin ainoa, jota tarkasteltiin yleisenä mallina suoraan uuteen järjestelmään ja joka myös täsmäkoulutettiin järjestelmään. Yleiset 6k- ja PVWatts-mallit antoivat samankaltaisia tuloksia, kuten myös täsmäkoulutetut 6k- ja PVUSA-mallit.

Yleisillä fysikaalisilla malleilla näytti olevan taipumusta mallintaa hetkittäinen teho lähes systemaattisesti liian suureksi. Tämä systemaattinen taipumus kasvatti energian tuotannon mallintamisen virhettä kuukausi- ja vuositasolla, kun mallinnetun tehon virheet kumuloituvat. Erityisen suuret energian tuotannon mallinnuksen virheet olivat Sodankylän järjestelmissä, jossa energian tuotanto mallinnettiin vuositasolla jopa 30 % liian suureksi. Kun fysikaalisia malleja täsmäkoulutettiin, systemaattinen taipumus mallintaa teho liian suureksi väheni, jolloin myös kuukausittaiset ja vuosittaiset energian tuotannon virheet pienenivät.

Työn toisena tutkimuskysymyksenä oli tutkia voidaanko koneoppimispohjaisilla malleilla vähentää tehon mallintamisen virhettä. Mallien kouluttamiseen käytettiin Helsingin, So-

dankylän ja Turun järjestelmien aineistoa ja luotujen mallien toimintaa testattiin Kuopion järjestelmässä. Näin saatiin kokeiltua miten hyvin mallit soveltuvat yleisiksi malleiksi joita voisi käyttää suoraan uuteen järjestelmään. Luodut mallit täsmäkoulutettiin myös osalla testiaineistosta. Tällä tarkasteltiin miten hyvin luotuja malleja voidaan muokata sopimaan paikalliseen järjestelmään.

Luoduista koneoppimismalleista yleinen HGBR-malli näytti R^2-arvon ja virhetermien mukaan sopivan paremmin uuteen järjestelmään kuin yleinen MLP-malli tai yleinen 6k-malli. Kun koneoppimismalleja täsmäkoulutettiin vuoden mittaisella testiaineiston pätkällä, MLP-mallin R^2-arvo kasvoi ja virhetermit pienenivät. HGBR-mallissa ei tapahtunut parannusta täsmäkoulutuksessa. HGBR-malli soveltui paremmin siis yleiseksi malliksi, kun taas MLP-malli oli parempi täsmäkoulutuken jälkeen. Yleinen HGBR-malli antoi suuremman R^2-arvon ja pienempiä virhetermejä, kuin yleiset fysikaaliset mallit testiaineistossa. Täsmäkoulutettu MLP-malli antoi puolestaan suuremman R^2-arvon ja pienempiä virhetermejä kuin täsmäkoulutetut fysikaaliset mallit.

Yleisillä koneoppimismalleilla kuukausittaiset ja vuosittaiset energian tuotannon virheet olivat pienempiä kuin yleisillä fysikaalisilla malleilla. Täsmäkoulutuksen jälkeen MLP-mallin kuukausittaiset ja vuosittaiset virheet yllättäen suurenivat, mutta HGBR-malleilla täsmäkoulutus ei juurikaan vaikuttanut kuukausittaisiin tai vuosittaisiin virheisiin.

Tuloksien perusteella näyttäisi siltä, että koneoppimismalleilla voidaan mallintaa PV-järjestelmän hetkellistä tehoa sekä energian tuotantoa kuukausi- ja vuositasolla Suomen oloissa tarkemmin kuin fysikaalisilla malleilla. Mallien tarkkuuksien erot eivät ole kuitenkaan suuria virhetermien perusteella. Merkittävänä erona mallien välillä näyttäisi olevan erityisesti yleisten mallien erilaiset systemaattiset taipumukset mallintaa tehoa. Yleiset fysikaaliset mallit mallintavat tehon systemaattisesti liian suureksi, kun taas yleiset koneoppimismallit mallintavat suuret tehon arvot liian pieniksi ja pienet tehon arvot liian suuriksi. Täsmäkoulutus vähentää systemaattisia taipumuksia sekä fysikaalisilla että koneoppimismalleilla.

Toinen merkittävä mallien välinen ero on siinä, että käytetyt koneoppimismallit ovat monimutkaisempia kuin fysikaaliset mallit. Koneoppimismallit ottavat syötteekseen 20 muuttujaa, mutta fysikaaliset mallit ottavat syötteekseen vain 3 tai 4 muuttujaa. Tämä voi hankaloittaa työssä käytettyjen koneoppimismallien käyttöä, sillä käytetyt muuttujat pitäisi saada mitattua tai laskettua. Syötemuuttujien lukumäärä voi tulla ongelmaksi erityisesti yksityishenkilöille tai aurinkosähkön pientuottajille, joilla ei ole kattavaa sääasemaa PV-järjestelmän yhteydessä. Toisaalta esimerkiksi kaupalliset tuottajat voisivat hyötyä mahdollisimman tarkoista malleista näiden mallien monimutkaisuudesta huolimatta.

LÄHTEET

[1] Fingrid, *Aurinkovoima*. url: https://www.fingrid.fi/sahkomarkkinainformaatio/aurinkovoima/ (viitattu 13. 03. 2025).

[2] Suomen uusiutuvat, *Toiminnassa olevat aurinkovoimalat*. url: https://suomenuusiutuvat.fi/aurinkovoima/aurinkovoimahankkeet-ja-voimalat-suomessa/toiminnassa-olevat-aurinkovoimalat/ (viitattu 17. 02. 2025).

[3] Suomen uusiutuvat, *Suunnittelussa olevat aurinkovoimahankkeet*. url: https://suomenuusiutuvat.fi / aurinkovoima / aurinkovoimahankkeet - ja - voimalat - suomessa / suunnittelussa - olevat-aurinkovoimahankkeet/ (viitattu 17. 02. 2025).

[4] C. Brester, V. Kallio-Myers, A. V. Lindfors, M. Kolehmainen ja H. Niska, ”Evaluating neural network models in site-specific solar PV forecasting using numerical weather prediction data and weather observations”, *Renewable Energy*, vol. 207, ss. 266–274, 2023. DOI: 10.1016/j.renene.2023.02.130.

[5] H. Böök, A. Poikonen, A. Aarva, T. Mielonen, M. R. A. Pitkänen ja A. V. Lindfors, ”Photovoltaic system modeling: A validation study at high latitudes with implementation of a novel DNI quality control method”, *Solar Energy*, vol. 204, ss. 316–329, 2020. DOI: 10.1016/j.solener.2020.04.068.

[6] L. Karttunen, S. Jouttijärvi, A. Poskela, H. Palonen, H. Huerta, M. Todorović, S. Ranta ja K. Miettunen, ”Comparing methods for the long-term performance assessment of bifacial photovoltaic modules in Nordic conditions”, *Renewable energy*, vol. 219, ss. 119 473–, 2023. DOI: 10.1016/j.renene.2023.119473.

[7] B. D. Dimd, S. Voller, U. Cali ja O.-M. Midtgard, ”A Review of Machine Learning-Based Photovoltaic Output Power Forecasting: Nordic Context”, *IEEE Access*, vol. 10, ss. 26 404–26 425, 2022. DOI: 10.1109/ACCESS.2022.3156942.

[8] A. Keddouda, R. Ihaddadene, A. Boukhari, A. Atia, M. Arıcı, N. Lebbihiat ja N. Ihaddadene, ”Solar photovoltaic power prediction using artificial neural network and multiple regression considering ambient and operating conditions”, *Energy Conversion and Management*, vol. 288, s. 117 186, 2023. DOI: 10.1016/j.enconman.2023.117186.

[9] M. B. Øgaard, H. N. Riise, H. Haug, S. Sartori ja J. H. Selj, ”Photovoltaic system monitoring for high latitude locations”, *Solar Energy*, vol. 207, ss. 1045–1054, 2020. DOI: 10.1016/j.solener.2020.07.043.

[10] S. Jouttijärvi, M. Tok, L. Karttunen, H. Huerta Medina, S. Ranta ja K. Miettunen, ”Modeling and Measuring the Power Output of Vertical Bifacial Solar Panels in Nor-

dic Conditions", *8th World Conference on Photovoltaic Energy Conversion; 507-511*, 2022. DOI: 10.4229/WCPEC-82022-3BO.14.2.

[11] N. Pearsall, *The performance of photovoltaic (PV) systems : modelling, measurement and assessment*, 1st edition. Amsterdam, Netherlands: Woodhead Publishing, 2017.

[12] IEC 61724-1, *Photovoltaic system performance - Part 1: Monitoring*, 2021.

[13] A. Luque ja S. Hegedus, *Handbook of photovoltaic science and engineering*, 2nd ed. Chichester, West Sussex, U.K: Wiley, 2011.

[14] W. Van Sark, A. Louwen, O. Tsafarakis ja P. Moraitis, "PV System Monitoring and Characterization", *Photovoltaic Solar Energy*, A. Reinders, P. Verlinden, W. Sark ja A. Freundlich, toim., 1. painos, Wiley, 2016, ss. 553–563. DOI: 10.1002/9781118927496.ch49.

[15] S. Daliento, A. Chouder, P. Guerriero, A. M. Pavan, A. Mellit, R. Moeini ja P. Tricoli, "Monitoring, Diagnosis, and Power Forecasting for Photovoltaic Fields: A Review", *International Journal of Photoenergy*, vol. 2017, nro 1, s. 13, 2017. DOI: 10.1155/2017/1356851.

[16] Sandia National Laboratories, *Plane of Array (POA) Irradiance*. url: https://pvpmc.sandia.gov/modeling-guide/1-weather-design-inputs/plane-of-array-poa-irradiance/ (viitattu 18. 10. 2024).

[17] Sandia National Laboratories, *Perez Sky Diffuse Model*. url: https://pvpmc.sandia.gov/modeling-guide/1-weather-design-inputs/plane-of-array-poa-irradiance/calculating-poa-irradiance/poa-sky-diffuse/perez-sky-diffuse-model/ (viitattu 30. 01. 2025).

[18] Y. Zhang, L. O. H. Wijeratne, S. Talebi ja D. J. Lary, "Machine Learning for Light Sensor Calibration", *Sensors*, vol. 21, nro 18, s. 6259, 2021. DOI: 10.3390/s21186259.

[19] J. Kratochvil, W. Boyson ja D. King, "Photovoltaic array performance model.", tekninen raportti SAND2004-3535, 919131, 2004. DOI: 10.2172/919131.

[20] V. Kallio-Myers, A. Riihelä, D. Schoenach, E. Gregow, T. Carlund ja A. V. Lindfors, "Comparison of irradiance forecasts from operational NWP model and satellite-based estimates over Fennoscandia", *Meteorological Applications*, vol. 29, nro 2, 2022. DOI: 10.1002/met.2051.

[21] B. R. Paudyal ja A. Gerd Imenes, "Performance assessment of field deployed multicrystalline PV modules in Nordic conditions", *2019 IEEE 46th Photovoltaic Specialists Conference (PVSC)*, Chicago, IL, USA: IEEE, 2019, ss. 1377–1383. DOI: 10.1109/PVSC40753.2019.8980629.

[22] R. E. Pawluk, Y. Chen ja Y. She, "Photovoltaic electricity generation loss due to snow – A literature review on influence factors, estimation, and mitigation", *Renewable and Sustainable Energy Reviews*, vol. 107, ss. 171–182, 2019. DOI: 10.1016/j.rser.2018.12.031.

[23] H. Böök, *Photovoltaic Output Modeling: Monitoring, Forecasting, and Applications*. Aalto University, 2021.

[24] H. Böök ja A. V. Lindfors, ”Site-specific adjustment of a NWP-based photovoltaic production forecast”, *Solar Energy*, vol. 211, ss. 779–788, 2020. DOI: 10.1016/j.solener.2020.10.024.

[25] H. Awad, M. Gül, K. M. E. Salim ja H. Yu, ”Predicting the energy production by solar photovoltaic systems in cold-climate regions”, *International Journal of Sustainable Energy*, vol. 37, nro 10, ss. 978–998, 2018. DOI: 10.1080/14786451.2017.1408622.

[26] A. Dolara, S. Leva ja G. Manzolini, ”Comparison of different physical models for PV power output prediction”, *Solar Energy*, vol. 119, ss. 83–99, 2015. DOI: 10.1016/j.solener.2015.06.017.

[27] T. Huld, G. Friesen, A. Skoczek, R. P. Kenny, T. Sample, M. Field ja E. D. Dunlop, ”A power-rating model for crystalline silicon PV modules”, *Solar Energy Materials and Solar Cells*, vol. 95, nro 12, ss. 3359–3369, 2011. DOI: 10.1016/j.solmat.2011.07.026.

[28] A. Dobos, ”PVWatts Version 5 Manual”, tekninen raportti NREL/TP-6A20-62641, 1158421, 2014. DOI: 10.2172/1158421.

[29] C. Whitaker, T. Townsend, J. Newmiller, D. King, W. Boyson, J. Kratochvil, D. Collier ja D. Osborn, ”Application and validation of a new PV performance characterization method”, *Conference Record of the Twenty Sixth IEEE Photovoltaic Specialists Conference - 1997*, 1997, ss. 1253–1256. DOI: 10.1109/PVSC.1997.654315.

[30] I. Goodfellow, Y. Bengio ja A. Courville, *Deep Learning*. MIT Press, 2016. url: https://www.deeplearningbook.org/ (viitattu 05. 09. 2024).

[31] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition*. O’Reilly Media, Inc, 2019.

[32] D. P. Kingma ja J. Ba, *Adam: A Method for Stochastic Optimization*, 2017. DOI: 10.48550/arXiv.1412.6980.

[33] D. Masters ja C. Luschi, *Revisiting Small Batch Training for Deep Neural Networks*, 2018. DOI: 10.48550/arXiv.1804.07612.

[34] Y. Bengio, *Practical recommendations for gradient-based training of deep architectures*, 2012. DOI: 10.48550/arXiv.1206.5533.

[35] T. J. Hastie, R. Tibshirani ja J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction* (Springer series in statistics), 2nd ed. New York: Springer, 2009.

[36] scikit-learn, *1.11. Ensembles: Gradient boosting, random forests, bagging, voting, stacking*. url: https://scikit-learn.org/stable/modules/ensemble.html (viitattu 27. 01. 2025).

[37] J. H. Friedman, ”Greedy Function Approximation: A Gradient Boosting Machine”, *The Annals of Statistics*, vol. 29, nro 5, ss. 1189–1232, 2001.

[38] B. Zazoum, ”Solar photovoltaic power prediction using different machine learning methods”, *Energy Reports*, 2021 The 8th International Conference on Power and Energy Systems Engineering, vol. 8, ss. 19–25, 2022. DOI: 10.1016/j.egyr.2021.11.183.

[39] I. Kayri ja M. T. Gencoglu, ”Predicting power production from a photovoltaic panel through artificial neural networks using atmospheric indicators”, *Neural Computing and Applications*, vol. 31, nro 8, ss. 3573–3586, 2019. DOI: 10.1007/s00521-017-3271-6.

[40] K. Wang, X. Qi ja H. Liu, ”A comparison of day-ahead photovoltaic power forecasting models based on deep learning neural network”, *Applied Energy*, vol. 251, s. 113 315, 2019. DOI: 10.1016/j.apenergy.2019.113315.

[41] M. J. Mayer ja G. Gróf, ”Extensive comparison of physical models for photovoltaic power forecasting”, *Applied Energy*, vol. 283, s. 116 239, 2021. DOI: 10.1016/j.apenergy.2020.116239.

[42] A. Pamain, P. V. Kanaka Rao ja F. N. Tilya, ”Prediction of photovoltaic power output based on different non-linear autoregressive artificial neural network algorithms”, *Global Energy Interconnection*, vol. 5, nro 2, ss. 226–235, 2022. DOI: 10.1016/j.gloei.2022.04.019.

[43] S. Khadke, B. Ramasubramanian, P. Paul, R. Lawaniya, S. Dawn, A. Chakraborty, B. Mandal, G. K. Dalapati, A. Kumar ja S. Ramakrishna, ”Predicting Active Solar Power with Machine Learning and Weather Data”, *Materials Circular Economy*, vol. 5, nro 1, s. 15, 2023. DOI: 10.1007/s42824-023-00087-5.

[44] A. Bouakkaz, A. Lahsasna, A. G. Mena, S. Haddad, M. L. Ferrari ja R. J. Castaneda, ”Evaluating and analyzing the performance of PV power output forecasting using different models of machine-learning techniques considering prediction accuracy”, *International Journal of Renewable Energy Development*, vol. 14, nro 1, ss. 158–167, 2025. DOI: 10.61435/ijred.2025.60547.

[45] D. Chicco, M. J. Warrens ja G. Jurman, ”The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation”, *PeerJ Computer Science*, vol. 7, s. 24, 2021. DOI: 10.7717/peerj-cs.623.

[46] S. Lindig, A. Louwen, D. Moser ja M. Topic, ”Outdoor PV System Monitoring—Input Data Quality, Data Imputation and Filtering Approaches”, *Energies*, vol. 13, nro 19, s. 5099, 2020. DOI: 10.3390/en13195099.

[47] A. Livera, M. Theristis, E. Koumpli, S. Theocharides, G. Makrides, J. Sutterlueti, J. S. Stein ja G. E. Georghiou, ”Data processing and quality verification for improved photovoltaic performance and reliability analytics”, *Progress in Photovoltaics: Research and Applications*, vol. 29, nro 2, ss. 143–158, 2021. DOI: 10.1002/pip.3349.

[48] M. Järvelä, K. Lappalainen ja S. Valkealahti, ”Characteristics of the cloud enhancement phenomenon and PV power plants”, *Solar Energy*, vol. 196, ss. 137–145, 2020. DOI: 10.1016/j.solener.2019.11.090.

[49] NREL, ”Evaluation of the Performance of the PVUSA Rating Methodology Applied to Dual Junction PV Technology: Preprint (Revised)”, 2009.

[50] scikit-learn, *Glossary of Common Terms and API Elements*. url: https://scikit-learn.org/stable/glossary.html (viitattu 10. 01. 2025).