Norges miljø- og biovitenskapelige universitet

Masteroppgave 2024 – 60stp
Fakultetet for Realfag og teknologi

Bruk av maskinlæring, lynobservasjoner og klimamodelldata til modellering av endring i lynaktivitet i Norge

Use of machine learning, lightning observation and climate model data to model change in lightning activity in Norway

Guro Størdal
Miljøfysikk og fornybar energi

Forord

Etter 6 år på Ås er det mange å takke. Alt fra de fantastiske folka jeg møtte i fadderuka som viste seg å bli venner for livet, til de man har møtt som frivillig én gang, og så er på hils med resten av studietida.

Det siste året har jeg vært fulltids masterstudent og kunne fordype meg i et utrolig spennende tema. Jeg valgte studiet mitt fordi jeg håpte på muligheten til å lære om klima og miljø, og det har virkelig toppet seg med denne oppgaven. Det har vært et slit til tider, og alle som har vært med underveis fortjener en takk. Veilederen min ved NMBU, Mareile Wolff, fortjener en stor takk for god støtte og oppmuntring. Takk for at du har hatt troa på meg hele veien. Alle ved Meteorologisk institutt har også vært utrolig viktige for meg med innspill, deling av data, kunnskap og erfaring og god hjelp når jeg spør. Spesielt takk til Anita, Andreas, Morten og Jean. Og også alle andre som har kommet med gode tips og innspill under kaffepauser og lunsj.

Å skrive master har vært utfordrende, og tidvis ensomt, men det har vært godt å ha folk ved siden av meg som skriver samtidig. Takk til gjengen som leverte høsten 2023, og den nye gjengen som leverer i mai 2024.

En spesiell takk må også rettes til Simen, for full oppvartning i innspurten av masteren, med ferdig middag hver gang jeg kommer hjem, og ingen oppvask å tenke på. Hunden min Dina fortjener også litt oppmerksomhet for god støtte og den viktigste bidragsyteren til at jeg har fått jevnlig frisk luft.

Takk for meg. Det er med skrekkblandet fryd jeg leverer masteren, og med det setter punktum ved studenttilværelsen på Ås.

Ås,
mai,
= 2024

Guro
Størdal

Sammendrag

Med et klima i endring er det mange værfenomener som kan endre karakteristikk og klimatologi. Et slikt værfenomen er lyn. I Norge er det lite forskning, men for andre deler av verden er det flere studier. De gir motstridende svar på om lynaktiviteten vil endre seg, både i mengde og lokasjon. Det er flere utfordringer med å modellere lynaktivitet for fremtiden. Den ene er tilgangen på gode nok data på nåværende og historisk lynaktivitet. En annen utfordring er hvordan man kan modellere lyn ut i fra andre parametere.

I denne oppgaven brukes simulerte klimadata fra tidligere studier gjort med HCLIM38-AROME, for en nåværende periode (2014-2018), en historisk periode (1986-2005) og en fremtidig periode (2081-2100) for månedene mai-oktober. Klimadataene fra den nåværende perioden brukes i kombinasjon med observerte lyn fra samme periode for å trene en maskinlæringsmodell for lynaktivitet. Parameterne som brukes tar utgangspunkt i hvilke parametere som brukes i værvarsling av lyn ved Meteorologisk institutt, men et stort utvalg modeller med kombinasjoner av flere parametere, der noen også tar i bruk funksjonsteknikk, prøves ut og sammenlignes. De som gir mest lovende resultater blir videreutviklet, og to modeller brukes til slutt for å modellere lynaktivitet for den historiske og den fremtidige perioden. Disse to periodene sammenlignes for å se om lynaktiviteten endrer seg.

Ingen av modellene som ble utprøvd ga et perfekt resultat, og det trengs mer forskning på hvordan bygge en god modell for lynaktivitet. Bedre og mer observasjonsdata er et punkt som raskt vil heve kvaliteten. To modeller med tilfeldig skog klassifisering (Random forest classifier, rfc) presterte bedre enn mange andre. Disse bruker nesten de samme parameterne, konvektivt tilgjengelig potensiell energi (Convective Available Potential Energy, CAPE), maksimal vertikal vindhastighet (w), 0-isotermen, en land/vann-maske, klokkeslett og konvektiv hemming (Convektiv inhibition, CIN) multiplisert med 0-isotermen, men den ene bruker i tillegg CIN og -15-isotermen.

De to modellene ga forskjellige resultater endring i lynaktivitet. Den ene modellerte en svak økning, mens den andre modellerte en svak reduksjon midlet over månedene mai-oktober. Begge modellerte at juli og august forblir ganske stabilt, men at spesielt i

starten av mai, september og oktober kan det bli en økning.

Summary

With a changing climate many weather phenomona may change characteristics and climatology. One such weather phenomena lightning. There is few studies ragarding Norway, but for other parts of the world there's several studies. They give conflicting results on whether the lightning activity will change, both in frequency and location. There are many challenges with modeling future lightning activity. One is acsess to good data on present and historical activity. An other challenge is how you model lightning from other parameters.

In this thesis simulated climate data from earlier studier with HCLIM38-AROME, for a present (2014-2018), historical (1986-2005) and future periode (2081-2100) for the months May-October is used. Climate data from the present periode is used in combination with lightningobservation from the same period to train a machine learning algorithms on lightning activity. The parameters used take base in what parameters that are used in lightning forecast at the Norwegian Meteorological Institute, but a broad selection of models with combinations of different parameters, also putting feature engineering to use, are trained and compared. The most promesing ones gets developed furtger, and in the end two models are used to model the lightning activity for the historical and the future periode. The periode gets compared to assess if the lightning activy changes.

None of the models gave a perfect result, and more reacarch needs to be but into how to build a good model for lightning activity. Improved observational data is a aspect that would easily rise the quality. Two RandomForest models performed better than many others. These use almost the same parameters, Convective Available Potential Energy (CAPE) maximal vertical vindspeed (w), 0-isotherm, a land/water-mask, time and Convective Inhibition (CIN), but one uses CIN and -15-isotherm in addition.

The two models gave different results ragarding the change in lightning activity. One modeld a weak increase, while the other modeled a weak decrease over the periode May through October. Both models no big changes for July and Augustm but especially in the beginning of May, September and October may see an increase.

Innhold

Figurer

Tabeller

1. Introduksjon

Med et endret klima følger det mange konsekvenser, og at klimaet endrer seg er det lite tvil om. Dette gjelder hele verden, og hvert område i verden blir påvirket på forskjellig måte (Calvin mfl., 2023). I følge rapporten «Klima i Norge, 2100» forventes det mer ekstremvær, høyere temperaturer, mer nedbør og høyere havnivå i Norge (Hanssen-Bauer mfl., 2015). Dette er de store trekkene, men det er et komplisert bilde som også består av mange flere spesifikke natur-, klima- og værfenomener. Et slikt fenomen er lyn og torden.

Lyn og torden er et av de mer intense værfenomenene som de fleste opplever i løpet av livet. Forskning på lyn startet allerede på 1700-tallet, men til å være noe mange opplever, er det kun i nyere tid man har satt seg inn i fysikken som ligger bak (Dwyer og Uman, 2014). I Norge har enkelthendelser de siste årene, (Rivrud, 2016, NRK, 2021), kanskje gjort mange oppmerksomme på at lyn kan være livsfarlig, men hva den største trusselen er avhenger av hvor du befinner deg (Holle, 2014). Holle (2014) har funnet at dødsfall og personskader relatert til lyn har blitt redusert i utviklede land, men at en stadig større og mer kompleks infrastruktur gjør samfunnet som en helhet mer utsatt for lyn.

Problematikk rundt kraftforsyningen er en bekymring i Norge. I studien gjort av Midtbø mfl. (2011), var betydning av et endret framtidig lynklima motivasjon bak studiet. Det samme gjaldt Køltzow mfl. (2018). De fant begge at lynaktiviteten muligens vil øke, men at det er store usikkerheter i hvor mye og hvor hen. Begge viste også hvordan lynklimatologien i Norge har vært de siste årene. Den har vært preget av store sesongforskjeller, med mye lyn på Østlandet om sommeren, men mest langs kysten av Vestlandet om vinteren.

Utenom disse studiene er det gjort lite forskning på endring av lynaktivitet for fremtiden spesifikt i Norge. Det finnes flere studier som fokuserer på større områder, der Norge indirekte blir inkludert. I de fleste av disse studiene tyder resultatene på en økning i lynaktivitet i Norge. Rädler mfl. (2019) fant at frekvensen med alvorlige lynhendelser vil øke over Europa, inkludert Norge, på grunn av økt ustabilitet i atmosfæren. I en annen

studie sammenlignet Finney mfl. (2018) to forskjellige parametriseringen for lyn, og fant at avhengig av hvilken parametriseringenr han brukte økte eller minket aktiviteten globalt. Ved å se på området over Norge viser begge projeksjonene en økning. Begge disse studiene har ganske lav horisontal oppløsning, henholdsvis 50 km og 250 km, og det er ikke sett på variasjoner gjennom året. Kahraman mfl. (2022) undersøkte endringer i lynaktivitet gjennom året med høy romlig oppløsning på 2,2 km. De fant store variasjoner i endring av lynaktivitet over Europa. Et utsnitt av Sørnorge dekkes av studien, og her ble det modellert en økning i aktivitet, spesielt på sensommeren. En studie som så på endringer i lynaktivitet i arktiske strøk fant at aktiviteten øker lineært med temperatur (Holzworth mfl., 2021).

Flere studier tyder altså på at lynaktiviteten i Norge vil øke. Siden de siste studiene spesifikt for Norge ble gjort, har nye modeller og høyere oppløst data kommet til, og det er interessant å se på det på nytt. Å se på hvordan lynaktiviteten i Norge endrer seg er hovedmotivasjonen til denne oppgaven. Dette gjøres ved bruk av klimadata simulert i sammenheng med NorCP-prosjektet, et nordisk samarbeid for å modellere klima i værvarsling. Simulerte klimadata fra tre tidsperioder brukes i denne oppgaven: En historisk periode (1986-2005), dagens periode (2014-2018) og en fremtidig periode (1981-2100). En utfordring med disse dataene er at lyn ikke finnes som et eget parameter. Den andre motivasjonen for denne oppgaven er derfor å utvikle et metode for å modellere lynaktivitet ut i fra klimadataene.

Det er mange tilnærmeringer for å modellere lynaktivitet ut i fra andre parametere. Tidligere har lynvarsling har vært basert på observasjon av vær og kunnskap om klimatologi, og kun vært mulig å varsle noen timer i forveien. Nye observasjonsmetoder, nye modelleringer, mulighet for kraftigere beregninger og mer data har utbedret varslingen av lynaktivitet, og alvorlige stormer, betraktelig (Johns, 1992).

På Meteorologisk institutt har det de siste årene blitt utviklet en modell basert på konvektivt tilgjengelig potensiell energi (Convective Available Potential Energy, CAPE), konvektiv hemming (Convektiv inhibition, CIN), makimal vertikal hastighet i luftsøylen (w), og relativ fuktighet ved 700hPa (relative humidity, RH). Modellen baserer seg på logistisk regresjon og beregner en sannsynlighet for lyn som brukes i værvarslinga (Køltzow, 2023).

Andre studier har sett på andre parametere og metoder for å beregne lynaktivitet. Kahraman mfl. (2022) fokuserte på IFLUX, som beregner sannsynligheten for lyn ut i fra parametere som beskriver hydrometeorene ved -15-isotermen og vertikal bevegelse, gitt at atmosfæren er over en gitt stabilitetsindeks. Midtbø mfl., 2011, Køltzow mfl., 2018 og Bright mfl., 2004 ser også på grenseverdier for indekser. Indeksene de ser på er for CAPE, løfteindeks (Lifting Index, LI), Temperatur (T) og nedbør (Bright mfl., 2004), LI,

K-index, Totals-totals-indeks og Showalter-indeks (forklaringer gitt i artikkelen, Midtbø mfl., 2011) og T, CAPE og nedbør (Køltzow mfl., 2018).

Noen studier ser på å beregne en frekvens for lyn ved hjelp av formler eller forskjellige regresjonsmodeller. Finney mfl., 2018 sammenligner den beregnede lynraten ved å bruke skytopp-høyden (Cloud-top height, CTH) eller IFLUX. Rädler mfl., 2019 bruker RH, LI og vindskjær i additive logistiske regresjonsmodeller, noe Battaglioli mfl., 2023 bygger videre på i sin studie.

Det er med andre ord mange metoder og parametere for å beregne lynaktivitet. En metode som har blitt tatt i bruk innenfor flere felt i nyere tid er maskinlæring. Den har også blitt tatt i bruk innenfor vær- og klimamodellering (Kashinath mfl., 2021), også lyn spesifikt (Geng mfl., 2021). Maskinlæring gir mulighet for å oppdage nye sammenhenger i datasett som ikke nødvendigvis egner seg for andre metoder- Sensornettverket for lyn i Norge har forandret seg mye over kort tid. Det har medført mange forbedringer, men betyr også at tidsserien er karakterisert av mange brudd og kun kortere perioder kan betraktes som homogen. En slik periode overlapper med dagens periode med den simulerte klimadataen. Disse to datasettene brukes derfor i denne oppgaven for å trene en maskinlæringsmodell, som deretter brukes til å modellere endringer i lynaktivitet.

2. Teori

I denne oppgaver brukes begreper og teori fra flere fagfelt, hovedsaklig geofysikk og datavitenskap, mer spesfikt maskinlæring.

I første og andre delkapittel er fokuset geofysikk. Første del går spesifikt inn på elektrifisering av skyer, mens andre del går inn på meteorologiske fenomener som er interessante å se på i sammenheng med lyn. Kilden til andre delkapittel er Physics and Chemistry of Clouds (Lamb og Verlinde, 2011), med mindre annet er spesifisert.

Tredje delkapittel går inn på maskinlæring, og hva som trengs for å forstå denne oppgaven. Modellene som brukes beskrives kort, sammen med verifiseringsmetoder og kalibrering. Her brukes Python Machine Learning (Raschka og Mirjalili, 2019) som kilde, med mindre annet er spesifisert.

2.1 Elektrifisering av partikler og skyer

Lyn kan defineres som «en veldig lang gnist,[...] på over 1km» (Dwyer og Uman, 2014), og er en elektrostatisk utladning i atmosfæren mellom to områder med forskjellig elektrisk ladning. Disse områdene dannes av forskjellig ladede partikler som separeres og holdes fra hverandre frem til ladningsforskjellen er såpass stor at den overvinner isoleringsbarrieren skapt av lufta. Avstanden mellom feltene med forskjellig ladning er typisk 5-10km, men kan være opptil 100km. Feltene kan være innad i samme sky, mellom skyer eller mellom en sky og bakken (Dwyer og Uman, 2014).

Det er mange teorier på hvordan de ladede partiklene blir til, og det er mye usikkerhet rundt dette. Det involver prosesser på både mikro og makro nivå, noe som gjør det vanskelig å studere i felt eller å gjenskape i laboratorieforsøk. Noen teorier som er ganske etablert er konvektiv oppladning, og ladning gjennom sammenstøt, både i og utenom eksisterende felt (Soula, 2012, Dwyer og Uman, 2014, Lamb og Verlinde, 2011).

2.1.1 Konvektiv oppladning

En mekanismen på makronivå er konvektiv oppladning, som oppstår i et eksisterende felt. Dette feltet kan komme av ionisering i øvre del av atmosfæren, i en region kalt elektrosfæren som er rundt 60km opp i atmosfæren. Her er det normalt en positiv, stabil ladning på 300kV i forhold til bakken, som hovedsaklig dannes av UV-stråling fra sola som ioniserer luftmolekyler. På grunn av den positive ladningen i elektrosfæren dannes det et elektrisk felt med jordoverflaten. Dette er sterkest ved bakken, og svakest øverst. Henholdsvis rundt 100 V/m og 4 V/m. Dette feltet er stabilt i områder hvor det ikke er storm, og kalles «fair-weather field», eller pent-vær-felt. Pent-vær-feltet opprettholdes ved at tordenskyer tilfører nye positive ladninger til elektrosfæren, og negative ladninger til jordoverflaten (Lamb og Verlinde, 2011).

Elektrifisering av skyer ved konvektiv oppladning starter med oppstrømmende luft som frakter med seg positivt ladede ioner på aerosoler fra grenselaget. Den positive ladningen kommer fra forskjellige kosmiske prosesser som skaper ioner som tiltrekkes av den negative jordoverflaten. Etter hvert som partiklene beveger seg opp i den frie troposfæren tiltrekker de seg negativt ladede ioner, og disse legger seg som et lag utenpå den voksende skya. Dette laget avkjøles når vannet fordamper, og fører til en synkende bevegelse. Når de negative ionene fraktes nedover forsterkes det elektriske feltet, og det dannes enda flere ioner som igjen fører til en forsterket strøm positive ioner inne i skyen og negative utenpå. Under helt ideele forhold kan dette føre til stor nok ladningsforskjell til å utløse lyn, men i praksis trengs også andre mekanismer, som ladning gjennom sammenstøt (Lamb og Verlinde, 2011, Soula, 2012).

2.1.2 Ladning gjennom sammenstøt

Ladning ved sammenstøt er en mekanisme på mikro-skala der partikler som kolliderer men ikke fester seg til hverandre, utveksler ladninger og blir forskjellig ladd. Typisk blir de største og tyngste, som regel graupel partikler, negativt ladet, mens de mindre og lettere, som regel snø- og iskrystaller, positivt ladet. De tyngste synker nedover relativt til de lette, og det blir en samlet positivt ladning øverst og en negativ nederst (Dwyer og Uman, 2014). Dette er riktignok ikke alltid tilfellet, og det kommer an på forholdene, som temperatur og relativ fuktighet, hvordan ladningene blir fordelt (Soula, 2012).

Ladning gjennom sammenstøt kan skje både i og uten et eksisterende elektrisk felt. I et eksisterende felt er partiklene polarisert, og vil tiltrekke seg ioner med motsatt ladning av hva de har på bunnen, og frastøte seg de som har lik ladning. I det elektriske pent-vær-feltet vil partiklene være orientert med positiv ladning nedover, og negativ ladning oppover. Dette vil si at de største tyngre partiklene vil tiltrekke seg negativt ladede ioner, og få en samlet negativ ladning, mens de positivt ladede ionene vil bli

frastøtt bunnen og hvis det er stor nok oppdrift, fortsette opp forbi den negativt ladede toppen. Dette fører til en samlet negativ ladning i bunnen av skyen, og en positiv lengre opp. For ladning gjennom sammenstøt der det ikke er et eksisterende felt er det mer omdiskutert hvordan ladningene blir fordelt, men hvordan de forskjellig type partiklene holder på ladningene kan ha en innvirkning (Lamb og Verlinde, 2011, Soula, 2012).

Kollisjonene skjer normalt i områder av skyen hvor temperaturen er rundt -10°C til -20°C (Dwyer og Uman, 2014). Etter at ladningene er overført må partiklene separeres for at de ikke skal gjenforenes og utligne hverandre. Her spiller størrelse på partiklene og hvor kraftig vertikal bevegelse i skya er, men et eksisterende elektrisk felt kan også være med på å separere partiklene (Lamb og Verlinde, 2011, Soula, 2012).

2.1.3 Struktur på en tordensky

Den vanlige strukturen på en moden tordensky er en stor overvekt av positiv ladning øverst i skyen, mens lengre nede er det negativ ladning. Dette kommer av hvordan partiklene typisk blir elektrifisert, som beskrevet i de foregående delkapitlene. Det er riktignok viktig å merke seg at strukturen på en tordensky kan være helt motsatt, eller med flere forskjellige ladningsfordelinger. Bakken kan også være både positivt og negativt ladet. Alt kommer an på hver enkelt lynhendelse, og hvilke mekanismer som har ledet til den (Lamb og Verlinde, 2011, Dwyer og Uman, 2014).

I en typisk lynhendelse vil det bevege seg negativt ladning, elektroner, fra skyen og til bakken. Men det kan også være positiv ladning fra bakken og opp, positiv ladning fra skya til en negativ overflate, eller negativ ladning fra bakken og opp til skya. Det kan også være utveksling av ladninger innad i en sky, mellom skyer, med lufta rundt eller elektrosfæren over (Dwyer og Uman, 2014).

2.1.4 Oppsummering av lyn-genererende prosesser

Grunnprinsippene for en lynhendelse kan oppsummert deles inn i 1) Separasjon av partikler og 2) Partikler som kan utveksle ladninger. For at 1) skal skje er man avhengig av en ustabil atmosfære slik at luft kan stige. For at 2) skal skje er man avhengig av hydrometeorer ved riktig temperatur. I de fleste klimamodeller finnes ikke «Ustabil atmosfære» og «hydrometeorer ved riktig temperatur» som parametere, men det finnes i stedet andre parametere som kan gi en god indikasjon. I neste delkapittel vil slike parametere diskuteres.

2.2 Parametere som kan indikere lynaktivitet

Av parameterne tilgjengelig i klimamodellen som brukes, er det utvalgte som er interessante med tanke lynaktivitet: Konvektivt tilgjengelig potensiell energi (Convective Available Potential Energy, CAPE, konvektiv hemming (Convektiv inhibition, CIN) og vertikal maksimal vindhastighet (w) forteller om stabiliteten i ei luftsøyle. Relativ fuktighet (Relative Humidity, RH), is-vann-bane (Ice water path, clivi), 0-isotermen og -15-isotermen sier noe om mengden hydrometeorer og hvilken fase de er i.

2.2.1 CAPE - Convective Available Potential Energy

Parameteret CAPE viser hvor mye energi som er tilgjengelig i en luftpakke når den løftes forbi nivået for fri konveksjon (Level of Free Convection, LFC). Når luftpakka løftes forbi LFC vil den ha positiv oppdrift, og akselerere oppover. Etter hvert som den avkjøles, vil vanndamp kondensere, latent varme frigjøres og oppdriften fortsette. Luftpakka fortsetter å stige frem til den når nivået for nøytral oppdrift (Level of Neutral Buoyancy, LNB). Her vil den ha samme tetthet som lufta rundt, og slutte å stige. Da vil mye av vanndamp være omgjort til vanndråper, og den potensielle energien i vanndampen er frigjort. En luftpakke med høy temperatur og mye vanndamp vil dermed ha en høyere CAPE og nå høyere opp, enn en tørr og kald en.

Lyn assosieres ofte med sterk oppdrift og konvektive skyer (Ávila mfl., 2010, Ushio mfl., 2001, Yoshida mfl., 2009). CAPE kan derfor egne seg som en indikator på lynaktivitet. En høy verdi på CAPE betyr mye potensiell energi og oppdrift, og dermed potensiale for lyn.

CAPE kan beregnes på mange måter, men dataene som brukes i denne oppgaven beregnes med ecmwf's metode (Groenemeijer mfl., 2019). Formelen for CAPE er som følger:

$$CAPE = \int_{LFC}^{LNB} B\mathrm{d}z \approx \int_{LFC}^{LNB} \frac{T_{v,parcel} - T_{v,env}}{T_{v,env}} g\mathrm{d}z \tag{2.1}$$

Hvor CAPE er «Convective Available Potential Energy», LNB er «Level of Neutral Buoyancy», LFC er «Level of Free Convection», og B er Bouyancy (Oppdrift) som integreres over høyden dz. Dette er tilsvarende det andre uttrykket der $T_{v,parcel}$ er den virtuelle temperaturen til luftpakka, $T_{v,env}$ er den virtuelle temperaturen til omgivelsene og g er gravitasjonskonstanten.

2.2.2 CIN - Convective Inhibition

Selv om det kan være høy CAPE i et område, kan luftpakken først utnytte denne energien hvis den løftes forbi LFC. CIN er en størrelse som forteller hvor mye energi som kreves for å gjøre dette. Er CIN høy, må det mye oppdrift til før luftpakka løftes til LFC, mens lavere CIN krever mindre oppdrift for å få konveksjon videre opp i atmosfæren.

CIN kan også beregnes på mange måter, men dataene som brukes i denne oppgaven beregnes med ecmwf's metode (Groenemeijer mfl., 2019). Formelen for CIN er som følger:

$$CIN = -\int_{surface}^{LFC} LFCB \mathrm{d}z \approx \int_{surface}^{LFC}\int_{LFC}^{LNB} \frac{T_{v,parcel} - T_{v,env}}{\overline{T}_v} g \mathrm{d}z \tag{2.2}$$

Hvor CIN er Convective Inhibition, surface er overflaten, og resten er tilsvarende som i formel 2.1.

2.2.3 W - Maksimal vertikal vindhastighet

Siden CAPE og CIN kun er potensielle parametere med stor usikkerhet (Groenemeijer mfl., 2019), kan andre parametere som også beskriver vertikal bevegelse og stabilitet inkluderes. Vertikal hastighet er et direkte mål på vertikalbevegelsen i en luftsøyle.

Vertikal hastighet varierer, så for å forenkle det sees det kun på maksimal vertikal hastighet i en luftsøyle. Vertikal vind har betydning for hvor mye luftmasser i ei luftsøyle forflytter seg, Yang mfl., 2016, og er derfor interessant med tanke på lyn og transport av hydrometeorer.

2.2.4 Relativ fuktighet

Relativ fuktighet (Relative Humidity, RH) defineres blant annet som forholdstallet, eller en prosent, mellom vanndampens partialtrykk i lufta, og vanndampens metningstrykk ved gitt temperatur. RH kan defineres som

$$RH = \frac{e}{e_s} \tag{2.3}$$

Hvor e er partialtrykket og e_s er metningstrykket. Ved overmetting kan RH ha en verdi på over 100% (Stull, 2017).

Klimadataene inneholder ikke RH eller partialtrykk, men temperatur, T, og spesifikk fuktighet, q. Spesifikk fuktighet defineres som

$$q \equiv \frac{\rho_v}{\rho_{air}} = \frac{n_v M_v}{n_d M_d + n_v M_v} \tag{2.4}$$

Hvor q er spesifikk fuktighet, ρ er tetthet, n er antall partikler, M er Molære masse, for henholdsvis vanndamp (subskript v) og tørr luft (supskript d). Molare masser er konstanter og forholdet $\frac{M_v}{M_d} = 0.622$. Det totale antallet partikler, n, vil i blander luft være summen av vanndamp- og luftmolekyler: $n = n_d + n_v$. Ved å dele det siste uttrykket i ligning 2.7 på M_d og substituere $n_d = n - n_v$ får man uttrykket

$$q_v = \frac{0.622 n_v}{n - 0.378 n_v} \tag{2.5}$$

Videre kan man flytte om og få

$$\frac{n_v}{n} = \frac{q}{0.622 + 0.378q} \tag{2.6}$$

Hvor q er spesifikk fuktighet. På bakgrunn av den ideelle gassloven, $p = nRT$ med den universelle gasskonstanten R kan man substituere n og n_v med henholdsvis p, totale trykk, og e, partialtrykk. Dette kan uttrykkes som:

$$e = \frac{qp}{622 + 0.378q} \tag{2.7}$$

Hvor e er partialtrykket, q er spesifikk fuktighet og p er det totale trykket.

I ligning 2.3 brukes også metningstrykk, e_s. Uttrykket for metningstrykket kan uttrykkes med Clausius-Clapeyron ligningen som

$$e_s(T) = e_0 e^{\frac{l_v}{R}\left(\frac{1}{T_0} - \frac{1}{T}\right)}, \tag{2.8}$$

hvor $e_s(T)$ er metningstrykket uttrykt ved temperatur, e_0 er en konstant, 611,3 Pa, l_v er spesifikk latent varme for vanndamp, R er den universelle gasskonstanten, T_0 er 0°C målt i kelvin, 273,15K, og T er temperaturen målt i kelvin. Forholdstallet $\frac{l_v}{R}$=5423K er konstant.

Ligning 2.7 og 2.8 kan substitueres inn i ligning 2.3, og man får da uttrykket,

$$RH = \frac{qp}{(0.622 + 0.378q)e_0} e^{-\frac{l_v}{R}\left[\frac{1}{T_0} - \frac{1}{T}\right]} \tag{2.9}$$

som gir en formel for RH uttrykt med spesifikk fuktighet, trykk og temperatur.

2.2.5 Isotermer

Isotermer er områder med samme temperatur, og kan markeres på trykkflater eller på høydelag. Standardtemperaturen i atmosfæren synker vanligvis fra jordoverflata og opp til tropospausen. Det er temperaturen i troposfæren som er relevant med tanke på lyn, siden tordenskyer sjeldent utvikler seg forbi tropospausen. Trykket synker også med høyde, og ved å legge en isoterm på høydedrag viser et lavere trykk at isotermen er høyere opp.

2.2.6 Clivi - Ice Water Path

Clivi beregnes ved å integrere isvann-innholdet i en sky, og beskriver mengden isvann per enhet areal. Andre studier har brukt lignende parametere i lynstudier, som is-blanding-ratio, (Geng mfl., 2021), oppover skyis-flux (Finney mfl., 2018) og nedbør (i sammenheng med CAPE) (Romps mfl., 2018).

2.3 Maskinlæring

Maskinlæring kan brukes som et effektivt verktøy til å se etter nye sammenheng og informasjon i store datasett. I denne oppgaven prøves to forskjellig type modeller ut, og det flere begreper og teori brukes for å måle ytelse og vurdere dem. I dette delkapittelet defineres og forklares det viktigste for å forstå, boka Raschka og Mirjalili (2019) brukes hovedsklig som kilde, med mindre annet er spesifisert.

I maskinlæringsmiljøet brukes ofte begrepet «features» for det som i andre fagmiljøer kalles parametere, variabler (eller indekser). I denne oppgaven brukes parametere.

2.3.1 Forskjellige maskinlæringsmodeller

Innenfor maskinlæring finnes det flere forskjellige modeller, men en av hovedgruppene, og hva som brukes i denne oppgaven, er overvåket læring (supervised learning). I overvåket læring blir modellene matet med data som er ferdig klassifisert, den kan utvikle seg hvis det blir tilført mer data, og man kan bruke den til å predikere verdier. Modellene som brukes i denne oppgaven er fra scikit-learn prosjektet (Pedregosa mfl., 2011).

Logistisk regresjon er en overvåket læringsmodell. Den er relativ enkel og egner seg veldig godt til lineære separerte data. Ved å sette en grenseverdi kan man klassifisere forskjellige hendelser, men man kan også få sannsynligheten for forskjellige hendelser. Dette gjør den godt egnet til blant annet værmeldinger.

Tilfeldig Skog Klassifiseringssystem (Random Forest Classifier, RFC) er en annen overvåket læringsmodell som baserer seg på flere beslutningstrær (Decision Trees). Et beslutningstre lærer ved å stille enten/eller spørsmål ved de forskjellige parameterne. For kategoriske parametere er det da enten én verdi eller én annen, men for kontinuerlige parametere vil treet sette en grense og skille mellom over og under denne grensen.

Et tre alene er utsatt for overtilpasning og bias, men med et ensemble unngår man dette. Et slikt ensemble med bestemmelsestrær kalles en tilfeldig skog. En tilfeldig skog har flere fordeler, og kan blant annet beregne hvilke parametere som gjennomsnittlig reduserer urenheter mest, og rangere dem etter det. En tilfeldig skog beregner sannsynligheten til hvilken klasse et datapunkt tilhører, og man kan få både sannsynlighetene og klassen med høyest sannsynlighet som et output.

Hyperparametere. Maskinlæringsmodeller har hyperparametere som regulerer læringsprosessen. Hver modelltype har flere hyperparametere, og forskjellige kombinasjoner av disse kan gi forskjellig utslag på hvor bra en modell presterer.

2.3.2 Preprosessering

Før en maskinlæringsmodell kan trenes må datasettet forberedes. Det omfatter utforskning av parametere og rensing av datasettet. I dette delkapittelet forklares teori og begreper som brukes i oppgaven for å utforske og preprosessere dataene.

Korrelasjonsmatrise visualiseres lineær korrelasjonen mellom kontinuerlige parametere. Korrelasjonen har en verdi mellom -1 og 1. En korrelasjon på 1 betyr at parametrene er korrelerte, da vil de øke og minke like mye, og motsatt for -1. Har de 0 korrelasjon er det ingen lineær sammenheng mellom dem. En korrelasjon nermest mulig 0 er en ønskelig.

Sannsynlighetstetthetsplot er en versjon av histogram hvor høyden av hver søyle er delt på totalt antall datapunkter (Hunter, 2007). Dette gjør at den er egnet til å sammenligne forelingen av parametere med forskjellig antall datapunkter.

Standardisering. For modeller som logistisk regresjon må parameterne standardiseres hvis de har forskjellig størrelsesorden. En formel er MinMax-scaler. Denne skalerer alle verdiene til et gitt område, for eksempel mellom 0 og 1. Høyeste verdi i parameterer blir tilsvarende 1, og laveste blir tilsvarende 0. Formelen er

$$X_{std} = \frac{X - X.min(axis = 0)}{X.max(axis = 0) - X.min(axis = 0)} \tag{2.10}$$

$$X_{scaled} = X_{std} * (max - min) + min$$

Hvor X er parametret som skal skaleres, $X.min(axis = 0)$ er minste verdi i dette parametret, $X.max(axis = 0)$ er største verdi i dette parametret, X_{std} er parametret standardisert, max er øvre grense av området, min er nedre grense av området og X_{scaled} er parametret ferdig skalert innenfor området (Buitinck mfl., 2013).

Funksjonsteknikk, på engelsk kalt «feature engineering», gir muligheten til å hente ny informasjon ved å kombinere eksisterende parametre. Dette kan frigjøre kapasitet i en maskinlæringsmodell, og gjør det mulig for den å gå dypere inn i andre parametere (Dong og Liu, 2018).

2.3.3 Ytelse og kalibrering

Det er mange ulike måter å måle ytelse og kalibrere modeller på. Her presenteres teorien bak de som brukes i denne oppgaven.

Oppdeling av datasettet. Ved trening av en maskinlæingsmodell deles dataene opp i tre deler. Trenings-dataene består av størsteparten av datasettet, og er hva som modellen trener på. Validerings-dataene er som regel et utdrag av trenings-dataene, og brukes av modellen for å sjekke hvor godt den presterer, og hvordan det eventuelt skal justeres. Test-dataene brukes etter at modellen er ferdig trent for å se hvordan modellen presterer på ukjent data som ikke har vært med i treningsprosessen. Hvis modellen overtilpasser seg til treningsdataene vil den være for spesifikk og ikke kunne gjøre gode predikasjoner med nye datasett. Hvis man vil bruke modellen til å predikere utfall for ny data vil man derfor ikke ha en modell som gjør det utelukkende bra på treningsdataen. Undertilpasning derimot er at modellen er for lite spesifikk og ikke klarer å finne mønstre i dataene og gir dårlige predikasjoner.

Ved **kryss-validering** blir samme datasett delt opp i flere trenings- og validerings-sett, og en modell trenes og valideres på alle disse settene etter tur. For hvert sett blir det produsert en score for hvor bra den presenteres, som kan brukes for en gjennomsnittlig score.

Binære klasser. I et binært datasett hvor man har to klasser er det fire begreper som er viktige å forstå:

- En sann positiv (True positive, TP) er et datapunkt som blir predikert til å være positiv, og som faktisk er det.
- En falsk positiv (False positive, FP) er et datapunkt som blir predikert til å være positivt, men som ikke er det.
- En sann negativ (True negative, TN) er et datapunkt som blir predikert til å være negativt, og som faktisk er det.

- En falsk negativ (False negative, FN) er et datapunkt som blir predikert til å være negativ, men som ikke er det.

Ytelsesmål for balanserte datasett. Nøyaktighet (accuracy, ACC) og feil (error, ERR) gir en enkel oversikt over hvor nøyaktig modellen er. Formlene for disse er henholdsvis

$$ACC = \frac{TP + TN}{FP + FN + TP + TN} \tag{2.11}$$

og

$$ERR = \frac{FP + FN}{FP + FN + TP + TN} \tag{2.12}$$

For et balansert datasett er ACC og ERR gode til å beskrive treffsikkerheten til modellen, men for et ubalansert datasett hvor andelen positive og negative hendelser ikke er 50/50, men kanskje 5/95 vil det gi et unøyaktig bilde. For et datasett der for eksempel 5% er positive og 95% er negative hendelser vil modellen oppnå 95% nøyaktighet ved å si at alle hendelser er negative. Derfor trengs det ander formler som kan gi et mer balansert bilde.

Ytelsesmål for ubalanserte datasett. Sann positiv rate (True positive rate, TPR) og falsk positiv rate (false positive rate, FPR) egner seg godt til ubalanserte datasett. I meteorologi kalles ofte TPR for hitrate (Hit rate) og FPR for falsk alarm (False alarm). Videre vil disse brukes sammen for å favne både meteorologi og maskinlæring. Formlene for TPR/hitrate og FPR/falsk alarm er henholdsvis

$$TPR/hitrate = \frac{TP}{P} = \frac{TP}{FN + TP} \tag{2.13}$$

og

$$FPR/falskalarm = \frac{FP}{N} = \frac{FP}{FP + TN}. \tag{2.14}$$

FRP/falsk alarm og TPR/hitrate egner seg veldig bra for å se på ubalanserte datasett. De gir også et godt bilde på om modellen over- eller underestimerer. Ideelt sett skal TPR/hitrate være lik 1 og FPR/falsk alarm være lik 0. Har man en høy TPR/hitrate og en høy FPR/falsk alarm har man få falske negative, så det vil si at alle som er positive blir klassifisert som positive, men med en høy FPR/falsk alarm betyr det at mange hendelser som er klassifisert som positive egentlig skal være negative. Modellen

overestimerer. Motsatt, med både lav TPR/hit rate og FPR/falsk alarm, blir hendelser som egentlig er positive klassifisert som negative, mens negative hendelser blir klassifisert som negative. Modellen underestimerer.

F1-score. En beregningmåte som prøver å balansere ut så man unngår både over og underestimering er F1-score. Det er en kombinasjon av tilbakekalling (recall, REC), som er lik TPR/hitrate og presisjon (precision, PRE). PRE = $\frac{TP}{TP+FP}$, og en høy verdi tyder på riktige predikasjoner, mens en lav verdi tyder på overestimering. Formelen for F1-score er

$$F1 = 2\frac{PRE \times REC}{PRE + REC} \tag{2.15}$$

Jo nærmere F1 er 1, jo bedre er modellen på å predikere korrekt, uavhengig av hvor balansert datasettet er.

ROC. For binære klassifiseringsproblemer hvor man får en sannsynlighet for de to klassene kan man bruke mottaker driftsegenskaper (receiver operating characteristics, ROC) kurver for å måle ytelsesevne. En ROC-kurve plotter forholdet mellom TPR/hitrate og FPT/falsk alarm for terskelverdier mellom 0 og 1. Ved terskelverdi 0 klassifisert alle hendelser som positive og både TPR/hitrate og FPR/falsk alarm er lik 0. For terskelverdi lik 1 klassifiseres alle hendelser som negative, og både TPR/hitrate og FPR/falsk alarm er lik 1. Diagonalen til ROC-kurven tilsvarer tilfeldig gjetning, og så lenge plottet er til ventre for dette er modellen bedre enn tilfeldig gjetning. Arealet under kurven (Area under curve, AUC), er mellom 0 og 1, og jo nærmere dette er lik 1, jo bedre presterer modellen.

Reliabilitets diagram brukes for å sjekke om en modell er velkalibrert. Datapunktene grupperes først i intervaller basert på beregnet sannsynlighet. Deretter beregnes frekvensen av positive hendelser innenfor hvert intervall. Frekvens plottes mot intervall, og man får en graf som viser hvor godt modellen predikerer sannsynlighet i forhold til observasjoner. I det laveste intervallet, hvor predikert sannsynlighet for en positiv hendelse er lav, burde frekvensen for observerte hendelser også være lav, og motsatt for høye intervaller. Ligger grafen under diagonalen overestimerer den, men hvis den ligger over overestimerer den (J, 2020).

Kvantiler. For modeller som over- eller underestimerer kan man bruke kvantiler for å beregne en grenseverdi som kalibrerer beregnede sannsynligheter mot observert frekvens per tidsenhet. For en fordeling med observerte frekvenser finner man gjennomsnittet, α. Dette kalles 2. kvantilen, og markerer hvor halvparten av verdiene er over, og halvparten under. Ved å beregne β=1-α kan man finne hvor man må sette grenseverdien til beregnet

sannsynlighet for å få samme gjennomsnittlig frekvens per tidsenhet som observerte frekvens per tidsenhet.

3. Data og metode

Data og metode består av flere deler. Først presenteres datasettene, og hvordan de har blitt preprosessert. Deretter presenteres hvordan metoden for å komme frem til en maskinlæringsmodell har vært, og til slutt presenteres metoden for å modellere endring i lynaktivitet for historisk og fremtidig periode, og hvordan man kan se på disse forskjellene.

3.1 Datasett

Det er to typer data som brukes i denne oppgaven. Det ene er simulerte klimadata, og det andre er lynobservasjoner. For de simulerte klimadataene er det tre perioder, historisk periode (1986-2005), fremtidig periode (2081-2100) og en «nåværende» periode (2014-2018). Lynobservasjonene brukes i sammenheng med den nåværende perioden for å bygge en lynmodell, mens den historiske og fremtidige perioden brukes for å se på endringer i lynaktivitet.

3.1.1 Klimadata

Flere datasett ble vurdert til oppgaven, men dataene som så ut til å egne seg best er fra simuleringer med HARMONIE-Climate (HCLIM), syklus 38. Harmonie er et samarbeid mellom flere land innen værvarsling og klimamodellering. Dette er en klimamodell som fokuserer på høy nok oppløsning til å eksplisitt tillate modellering av konveksjon. AROME egner seg spesielt godt til regionale simuleringer på grunn av den høye oppløsninga (Belušić mfl., 2020), og er hva som har blitt brukt for å simulere klimadataene i denne oppgaven. De har en horisontal oppløsningen på 3km, med 3-timers tidsintervall. De beskriver nåverdier ved hoved og mellomliggende synoptiske tidspunkter.

Klimadataene er produsert som en del av NorCP-prosjektet, et nordisk samarbeid innen konveksjonstillatende klimamodellering. Klimadata for perioden 1998-2018 ble undersøkt i en studie av Lind mfl. (2020). De så på fordelene med klimamodeller som har høy nok oppløsning til å simulere konveksjon. Her er ERA-Interim brukt

som lateral grense. Studien viste at simuleringene stort sett stemte godt overens med observasjoner, og dette brukes derfor som «ground truth» i denne oppgaven. Siden den simulerte tidsperioden overlapper med tidsperioden for lynobservasjoner brukes disse to datasettene sammen for å bygge en lynmodell vha maskinlæring.

Endring i lynaktivitet analyseres ved å se på forskjellen mellom en historisk periode (1985-2005) og en fremtidig periode (2081-2100). Her er laterale grenser fra EC-EARTH med RCP8.5 som utslippsscenario. RCP8.5 er et høyutslippscenario, der utslippene fortsetter å øke, og er hva IPCC anser som det mest dramatiske scenarioet med størst konsekvenser (Calvin mfl., 2023). Klimadataene fra disse periodene er analysert i en studie gjort av Lind mfl. (2023). De fant at man kan forvente flere endringer i klima over Fennoskandina, spesielt med tanke på nedbør. De viste også at det er en stor fordel med modeller med høy nok oppløsning som tillater konveksjon når man skal modellere klima over denne regionen.

3.1.2 Lyndata

Lynobservasjonene er fra et nettverk av sensorer Meteorologisk institutt nå drifter. Frem til 2017 var det driftet og eid av Statnett, og har siden oppstart på 90-tallet blitt utbedret og rekalibrert i flere omganger. Det foreligger ikke dokumentasjon på alle disse endringene, noe som betyr at det er store usikkerheter i hvor sammenlignbare observasjonene er over lengre perioder. Noe som er kjent er at fra og med 2014 ble sky-sky lyn registrert i tillegg til sky-bakke lyn. De to lyntypene blir kategorisert blant annet ved grenseverdier som kan settes manuelt, og disse verdiene ble i årene etter 2014 justert flere ganger. I slutten av 2018 ble flere sensorer byttet ut. På grunn av dette omfattende sensorskifte er dataene fra før og etter slutten av 2018 ikke direkte sammenlignbar (Salomonsen, 2024, Sidselrud, 2024).

Lynobservasjonene som brukes i denne oppgaven er derfor satt til tidsperioden fra 2014 til 2018. I datasettet er begge lyntypene inkludert som en kategori: Lyn. I tidsrommet var det stor variasjon i hvordan disse ble klassifisert. Denne tidsperioden overlapper også med tilgjengelige klimadata som går fra 1998-2018, men begrenses til 2014-2018 fordi sky-sky ble inkludert først fra 2014.

Nettverket består av sensorer plassert rundt om i Norden, se figur 3.1. Her ser man hvor sensorene er plassert. De som er markert som røde var ikke i drift da figuren ble laget. Sensorene fanger opp det elektromagnetiske signalet til individuelle lyn. Dette signalet filtreres for støy og sendes til en lokal prosessor som beregner lokasjonen til lyn-nedslaget og hvilken type lyn det er. De tekniske spesifikasjonene kan leses om i brukermanualen (Oyj, 2015).

Figur 3.1: Kart over nettverk av lynsensorer i Norden, mai 2018. Grønne sensorer var i drift, mens røde var ute av drift da kartet ble produsert i 2018. Brukt med tillatelse fra Meteorologisk Institutt.

Produsenten av sensorene angir over 90% deteksjonevne for sky-bakke, og 50% for sky-sky under ideelle forhold. Under reelle forhold kan man derfor anta at deteksjonsevnen er noe lavere. En faktor er at Norge har en varierende topografi, og at det er store avstander mellom flere av sensorene. Dette gjelder spesielt på Vestlandet, hvor lokale tordenstormer i dype fjorden og daler ikke er like lett å registrere. I Nordnorge er det også store avstander, noe som gjør signalene vanskeligere å oppfatte. Et mer realistisk tall er at 60-70% av lyn registreres, og at det er spesielt mye lyn som ikke registreres på Vestlandet (Salomonsen, 2024).

Datasettet med lynobservasjoner var prosessert ved Meteorologisk institutt, og inneholdt summen av lynobservasjoner for hver halvtime før og etter hver hele time. Det vil si at dataene fra for eksempel 3. juni 2017 kl. 12:00 er summen av lyn mellom 11:30 og 12:30 den dagen, kl. 17:00 er for 16:30 til 17:30 osv. Ettersom parameterne kun er for hver tredje time måtte lyn dataene filtreres. For å kunne brukes sammen med kilmamodellparameterne som kun finnes hver trede time, ble lyndatasett redusert og kun data for de synoptiske tidspunktene ble brukt videre. Datasettet inneholdt to parametere for lynobservasjoner. Den ene beskriver antall lynobservasjoner for ei grid-rute innenfor gitt tidsrom. Den andre beskriver antall lynobservasjoner innenfor en radius på 25km. Siden målet med oppgaven er å lage en modell som kan gjenkjenne flest mulig værsituasjoner karakterisert av lyn, ble parameteret med antall lynobservasjoner innenfor 25 km radius brukt i denne oppgaven. Lynobservasjonene er på samme grid som The MetCoOp Ensemble Prediction System (MEPS), og har 2.5km oppløsning (Køltzow, 2023).

3.2 Preprosessering og oppbygging av datasettet

Før den egentlige analysen av dataene kunne begynne var det nødvendig med en del forarbeid. For klima- og lyndataene var en stor del av oppgaven å finne hvilke parametre som var nyttige, hvordan preprosessere dem, og hvordan redusere dem til en håndterbar størrelse.

3.2.1 Utforskning av parameterne

For å undersøke dataene ble det benyttet flere verktøy. For å undersøke korrelasjon mellom parameterne ble det beregnet en korrelasjonsmatrise, og parameterne ble vurdert etter hvor mye de korrelerte med hverandre.

For å se på fordelingen av lynobservasjoner mot verdiene til de forskjellige parameterne ble det brukt sannsynlighetstetthetsplot. Hele datasettet ble plottet sammen med delen av datasettet som inneholdt lyn, og det ga en indikasjon på hvilke intervaller av parameterne det var flere lynobservasjoner. Dette var også nyttig for å se hvilke parametere som tydelig separerte lynobservasjoner fra resten av datasettet.

Ved å multiplisere parameterne med hverandre ble funksjonteknikk brukt for å se etter nye mønstre. Før parameterne ble multiplisert ble de skalert med MinMax-scaler for å være i samme størrelsesorden. Resultatene fra funksjonsteknikken ble plottet i sannsynlighetstetthetsplot, og det ble vurdert om noen av plottene skilte lynobservasjoner fra resten av datasettet tydeligere enn om parameterne ble plottet alene.

Lyndataene ble også utforsket, og det ble sett på gjennomsnittlig antall lyn per tidsenhet, antall lynhendelser per tidsenhet, og når på året og døgnet det er flest tilfeller av lyn.

3.2.2 Preprosessering med CDO og gridpp

I figur 3.2 vises et flytdiagram av første del av preprosesseringa, der hovedsaklig verktøyene CDO og gridpp (utviklet av og for meteorologisk institutt Køltzow, 2023) ble brukt.

I flytdiagrammet står P for Parametre, L for Lyndata, M for Maske og D for Datasett for å lettere kunne referere til de forskjellige stegene av prosessen. Datasettet som ble brukt til trening av maskinlæringsmodellene fulgte hele prosessen, mens dataene brukt i modellering fulgte P1 til P3, og deretter D1 og D2 uten å bli sammensatt med lynobservasjonsdata ettersom det er disse som skal modelleres. De forskjellig stegene innbar som følger:

P1: Beregning av parametere. Ettersom ikke alle parametere som var av interesse eksisterte i datasettet måtte disse beregnes. 0- og -15-isotermene ble beregnet med interpolering mellom temperaturer på forskjellige trykknivåer. W ble filtrert ut fra vindhastigheter ved forskjellig trykknivåer, og RH ble beregnet med ligning 2.9 ut i fra spesifikk fuktighet og temperatur ved 700hPa. Resterende parametere trengte ikke beregnes.

P2: NorCP data. I dette steget ble all data for NorCP samlet i ei fil. Disse parameterne var CAPE, CIN, RH, w, trykknivå ved 0-isotermen og -15-isotermen, samt om et gridpunkt over vann eller land. Deretter ble alle NaN-verdier erstattet. Dette gjaldt CIN og isotermene. Nan-verdier for CIN ble erstattet med 900 Jkg^{-1}, ettersom dette er øvre grense for CIN-verdiene når CIN beregnes med klimamodellen. Isotermene har NaN-verdier der temperaturen ved overflaten er lavere enn isotermen. Ettersom isotermene blir beskrevet ut i fra trykknivå, ble NaN-verdiene erstattet med 150000 Pa siden dette er noe høyere enn bakketrykket. Det er minimalt med tilfeller av lyn når isotermene er NaN, og tanken var at en maskinlæringsmodell kan dra nytte av at høyt trykk betyr lite sannsynlighet for lyn. Det samme gjelder i stor grad når CIN = 900 Jkg^{-1}.

P3: NorCP u/NaN-verdier. I dette steget har man et datasett med all klimadata uten NaN-verdier. Videre ble en nabolagsmetode med persentiler fra gridpp brukt for å glatte ut CAPE, CIN, RH og w. Gridpp er et verktøy utviklet av og for Senteret for Utvikling av Værvarslingstjenesten (SUV) ved meteorologisk institutt, og brukes i produksjon av værmelding (Køltzow, 2023, MET, 2024).

L1: Lynobservasjoner hver time. Lynobservasjonene var opprinnelig med én times oppløsning, men ble her filtrer til hver 3. time for å kunne brukes sammen med klimadataene.

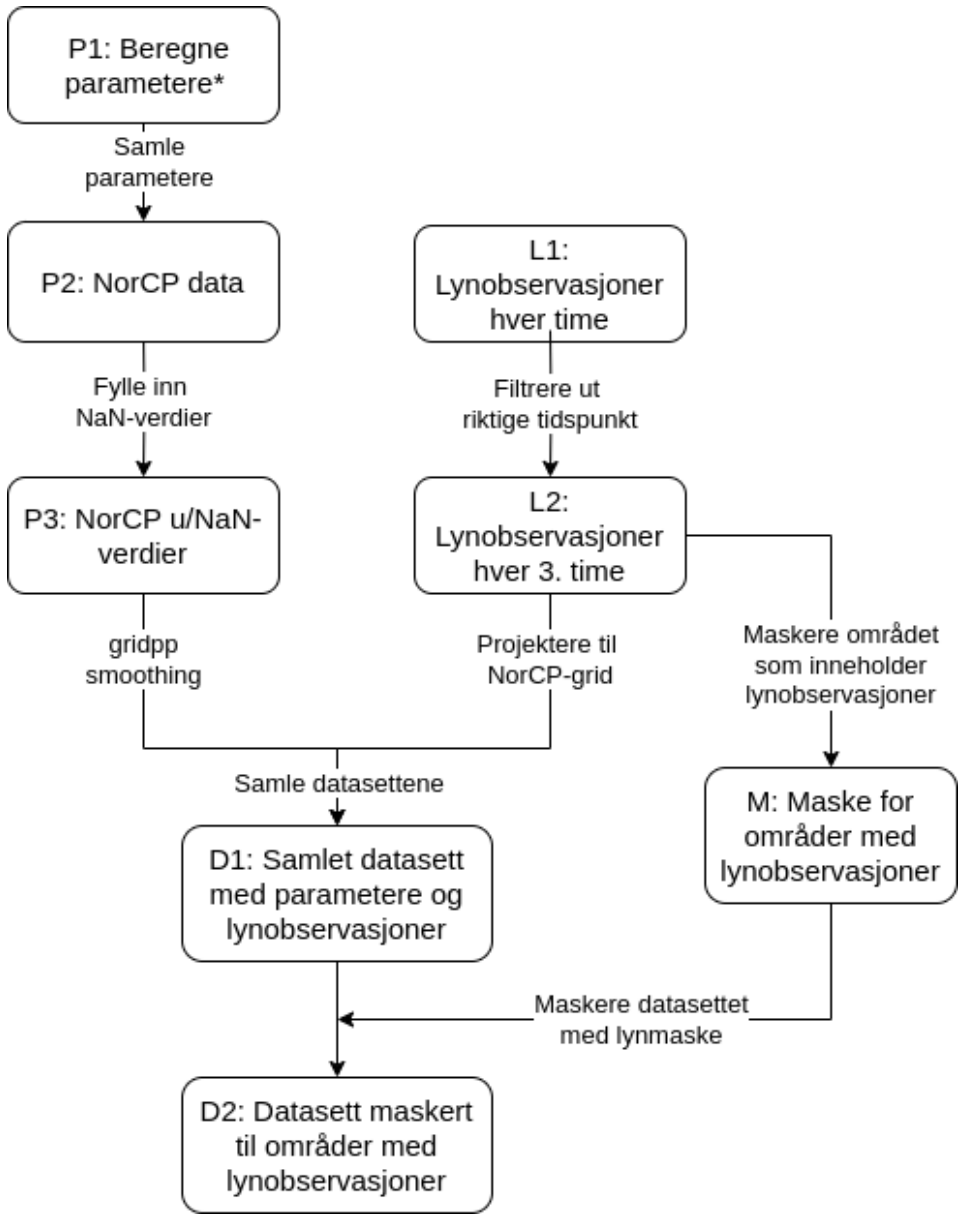
L2: Lynobservasjoner hver 3. time. For å kunne koble lynobservasjonene til parametere på samme lokasjon ble lyndataene prosjektert til samme grid som parameterne.

D1: Samlet datasett med parametere og lynobservasjoner. Parameterne og lynobservasjonene samles i et samlet datasett. Disse blir maskert med maske for områder med lynobservasjoner som forklares i neste trinn.

M: Maske for områder med lynobservasjoner. Området hvor det blir gjort lynobservasjoner dekker ikke hele det samme området som parametrene. Områdene utenfor dette blir ikke sett på i denne oppgaven, og må derfor fjernes for å lage et rent datasett. Masken ble laget ved at alle gridpunkter hvor det har blitt gjort en lynobservasjon innenfor 25km radius ble en del om datasettet. De resterende blir ikke

tilgjengelige.

D2: Datasett maskert til områder med lynobservasjoner. Dette datasettet er resultatet av de prosessene som blir gjort med CDO og andre verktøy i Linux-terminalen. Det består av alle parametere fra NorCP/HCLIM som skal prøves ut, enkelte **smoothed** med gridpp, og lynobservasjonene for hver tredje tilsvarende time. Videre preprosessering ble gjort med Python.



Figur 3.2: Flytdiagram for første del av preprosessering av datane, gjort med CDO og griddpp tilgjengelige via meteorologisk institutt. Forkortelser: P: Parametre, L: Lyndata, M: Maske og D: Datasett. Hver rute representerer et stadium på datasettet, mens pilene forklarer endringer som ble gjort. Flytdiagrammet er laget vha. draw.io **Enkelte beregninger er utført av kolleger fra Meteorologisk Institutt, deriblant beregning av 0- og -15-isotermene, og maksimal vertikal vindhastighet.*

3.2.3 Preprosessering i Python

Etter preprosessering i terminalen måtte det gjøres videre preprosessering som gjorde datasettet egnet til å jobbe med i Python. Det ble også redusert i størrelse for å bli mer håndterlig. Spesielt pakkene matplotlib (Hunter, 2007), pandas (McKinney, 2010), numpy (Harris mfl., 2020) og scikit-learn (Pedregosa mfl., 2011) har vært mye brukt.

Redusering av størrelse. For å redusere størrelsen ble datasettet redusert i både tid og rom. Denne studien begrenser seg til utvikling av lyn i Norge. Derfor ble et kartutsnitt for å fjerne overflødige områder utenfor interesse gjort. Videre ble kun tidsperioden mai til oktober videre med i analysen. Denne reduseringen ble anvendt for både treningsdata og dataene til modellering av historisk og fremtidig periode.

For treningsdataene ble i tillegg flere av datapunktene hvor CIN = 900 Jkg^{-1} fjernet. I ca 80% av datasettet var CIN=900 Jkg^{-1} og i mesteparten av disse ble ikke lyn registrert. Derfor ble 75% av tidspunktene av trengingsdatasett, der CIN=900Jkg^{-1} og ingen lyn var observert, fjernet.

Dataformat. Det foretrukne dataformatet for maskinlæringsmodellene som brukes i denne oppgaven er DataFrame. Dataene ble derfor omskrevet til DataFrames, og lagret som egne filer ved hjelp av modulen pickle. Dette ble gjort for all data.

Nytt parameter. Det ble også lagt til et nytt parameter, tidspunkt, som inneholdt tidspunktet for datapunktet. Dette ble hentet fra indeksen til datapunktet. For treningsdataene ble indeksen etter dette flatet ut for å redusere størrelse, men for modelleringsdatane ble indeksen videreført for å lettere kunne bygge det tilbake til en netcdf-fil.

3.3 Bygging av en maskinlæringsmodell

Med et ferdig preprosessert datasett var det mulig å begynne å prøve ut forskjellige maskinlæringsmodeller.

3.3.1 Modeller

En grunnmodell, logreg_0, var utgangspunktet. Den var basert på den operasjonelle lynalgoritmen ved Meteorologisk institutt. En beskrivelse av denne ligger som vedlegg A. Modell logreg_0 brukte de samme parameterne, CAPE, CIN, w og RH glattet ut med gridpp, men de ble skalert med MinMax-scaler mellom 0 og 1. Det var også kun w og RH som trengtes å beregnes, og tidsoppløsningen var på 3 timer. Disse dataene ble brukt i en logaritmisk regresjonsalgoritme for å få logreg_0.

Som et alternativ til logistisk regresjon ble også flere rfc-modeller prøvd. Det ble prøvd ut kombinasjoner av forskjellige parametere, blant annet noen av dem multiplisert med hverandre.

Dataene brukt til trening. Siden datasettet var stort, selv etter preprosessering og reduksjon, var det kun deler av datasettet som ble brukt til å prøve ut forskjellige konfigurasjoner. Alle modeller ble først trent på data fra 2015, siden dette året lå omtrent i midten på lynaktivitet. De modellene som så mest lovende ut ble trent på alle årene. Det ble også gjort videre reduksjon av datasettet før trening. For 2015 inneholdt datasettet 1.23% lynobservasjoner, men datapunkter med ikke-lyn ble tilfeldig redusert med pandas.DataFrame.sample(random_state = 1) slik at det var 5.45% lynobservasjoner. For 2014-2018 besto datasettet av 2.63% lynobservasjoner. Ved å redusere antall ikke-lyn med samme random_state ble andelen lyn økt til 6.18%.

I det opprinnelige datasettet er antall lynobservasjoner kvantifisert, med opptelling av antall lyn i hvert datapunkt. Dette varierte mellom 0 og 4207, men ble kategorisert til lyn/ikke-lyn siden spennet var så stort og det forenklet oppgaven betraktelig å gjøre lynobservasjonene til et binært parameter.

Pipeline og GridSeachCV. Skalering, funksjonsteknikk og modell-type ble bygd inn i pipelines for å redusere antall steg for å trene hver modell. Pipelinene ble deretter kjørt i et GridSearchCV sammen med et spekter av hyperparametere. GridSearchCV brukes for å tune hyperparameterne til modellene for å finne den beste kombinasjonen. Dette gjøres ved at alle kombinasjner av hyperparameterne prøves ut. Når alle kombinasjonene er testet ut og evaluert, blir den kombinasjonen som gir høyest score valgt som den beste scoren. Til slutt kan hele datasettet bli brukt til å trene modellen med de beste hyper-parameterne, og er hva GridSearchCV presenterer som den beste modellen. En fordel med GridSearchCV er at kryss-validering er innebygd, og man får en mer robust modell.

3.3.2 Utvelgelse av modeller

For å evaluere hvilke modeller som presterte best ble de vurdert opp i mot forskjellige ytelsesmål.

Ytelsesmålene som ble brukt var f1-score, ROC-AUC, reliabilitetsdiagram og visualisering av predikasjoner og sammenligning med observerte data.

For å kalibrere grenseverdien for lyn ble det brukt kvantiler. Den gjennomsnittlige frekvensen per tidsenhet for lynobservasjoner over land og vann ble beregnet, og ut i fra disse ble en grenseverdi for land og en for vann beregnet.

Med de nye grenseverdiene ble frekvensen per tidsenhet for hvert år beregnet og de modellene som klarte å rangere årene riktig fra høyest til lavest frekvens per tidsenhet vurdert som mer presise.

De modellene som viste mest lovende resultater ble trent på alle årene, og de samme ytelsesmålene sammenlignet. Deretter ble de modellene som hadde best resultater trent på all data. Det ble beregnet endelige grenseverdier, og modellene og grenseverdiene ble brukt til å predikere lynaktivitet for historisk og fremtidig periode.

3.4 Endring av lynaktivitet

For å se på endringer i lynaktivitet ble CDO brukt til beregninger. Årsgjennomsnittet, månedsgjennomsnittet og det daglige gjennomsnittet for frekvensen av lyn per tidsenhet ble beregnet får både den historiske og den fremtidige perioden. For å få et bilde av endringer i klimatologi trengs gjennomsnittet over lengre perioder, og et og et år er derfor ikke representativt når man ser på klimatologi (Lind mfl., 2023).

For å se på endringer ble fremtidig periode trekt fra den historiske, og prosent endring ble beregnet i python for både år, måned og dag. Disse resultatene ble visualisert og diskuteres i resultater og diskusjon.

3.5 Tekniske deltaljer

Dette delkapittelet opplyser om tekniske detaljer rundt metoden i oppgaven.

3.5.1 Bruk av kunstig intelligens

I denne oppgaven er den eneste formen for kunstig intelligens som er tatt i bruk maskinlæringsalgoritmene fra scikit-learn. Språkmodeller som ChatCTP og lignende er bevisst ikke benyttet (RealTek, 2024).

3.5.2 GitHub repository

All kode brukt i Python ligger tilgjengelig på GitHub. For interessert kan man se den på: https://github.com/gurosto/lynogtorden.git

4. Resultater

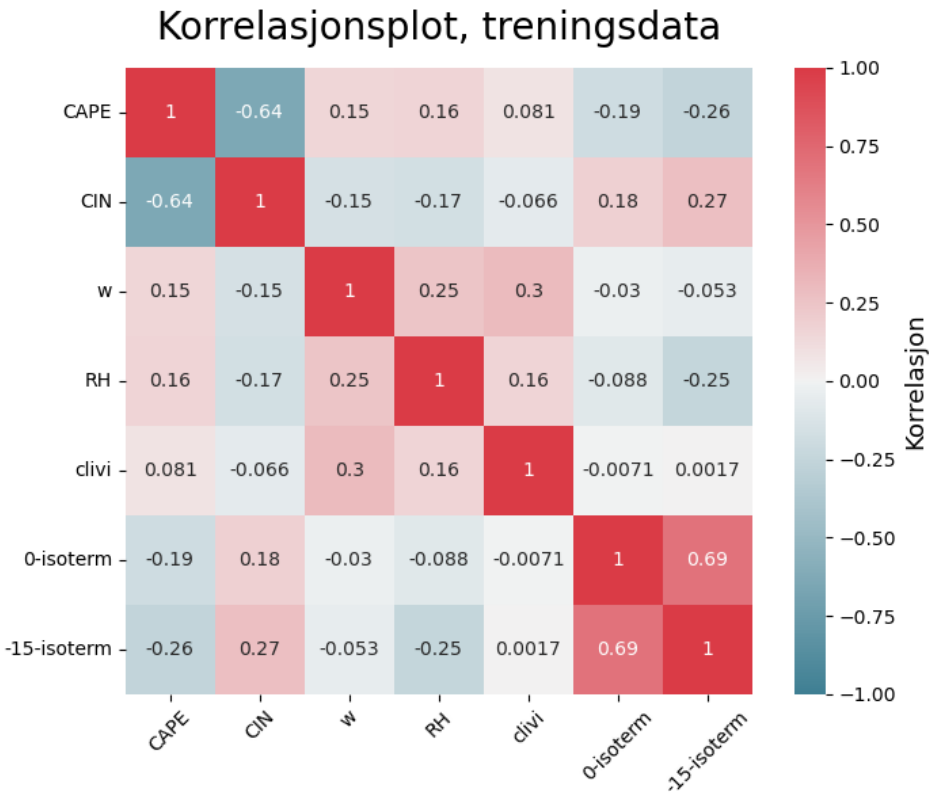
Resultatene i denne oppgaven starter med resultatene fra utforskningen fra data. Her blir forskjellige plot presentert og beskrevet, samt noen kart til lyndataene. Videre blir resultater fra maskinlæringsmodellene presentert. Dette er hovedsaklig mål på presisjon og ytelse, og plot til noen utvalgte modeller. Alle plot ligger som vedlegg. Til slutt vil endringene i lynaktivitet og parametere bli presentert. Dette gjøres med kart og plot som viser endring i aktivitet.

4.1 Utforskning av data

For å se hva jeg jobbet med startet jeg med å visualisere klimadataene fra mai-oktober for 2014-2018, og hente ut diverse info fra dem. Parameterne som ble utforsket var CAPE, CIN, w, RH, 0-isoterm, -15-isoterm og clivi.

4.1.1 Korrelasjonsmatrise.

I figur 4.1 er en korrelasjonsmatrise over de kontinuerlige parameterne brukt i treningen av maskinlæringsmodellen. De med sterkest rød farge er mest lineært korrelert, og de med mørke blå farge er mest negativt lineært korrelert. Jo lysere fargen er, jo mindre korrelert er de. Isotermene har den høyeste korrelasjonen med hverandre på 0.69, og CAPE og CIN har høyest negativ korrelasjon på -0.64. Ellers ligger de andre verdiene mellom -0.26 og 0.3. Clivi og -15-isotermen er minst korrelerte med en verdi på 0.0017.

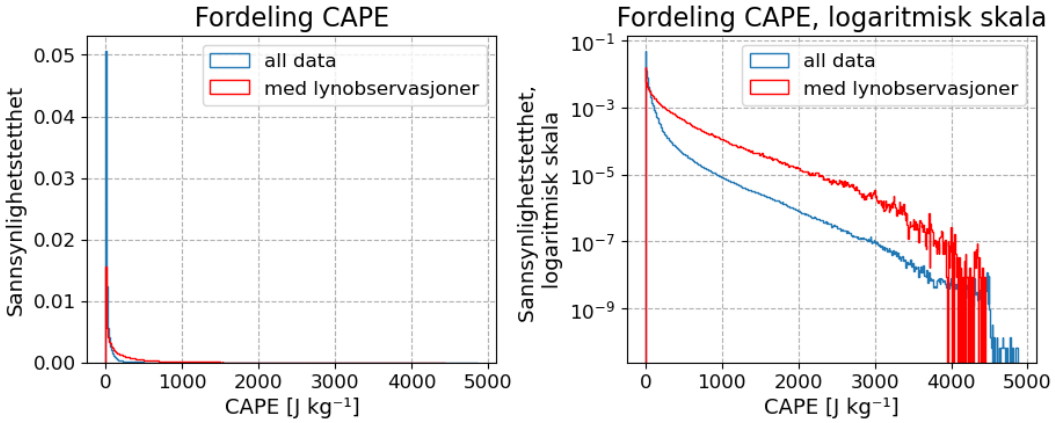


Figur 4.1: Korrelasjonsmatrise for de kontinuerlige parameterne i treningsdataene. Jo mørkere rød farge, jo nermere 1 og mer positivt lineært korrelerte er parameterne. Jo mørke blå farge, jo nermere 0 mer negativt lineært korrelerte er de. Lys farge betyr nerme 0 lav korrelasjon. Parameterne som er med er CAPE, CIN, w, RH, clivi, 0-isotermen og -15-isotermen.

4.1.2 Sannsynlighetstettshetplot

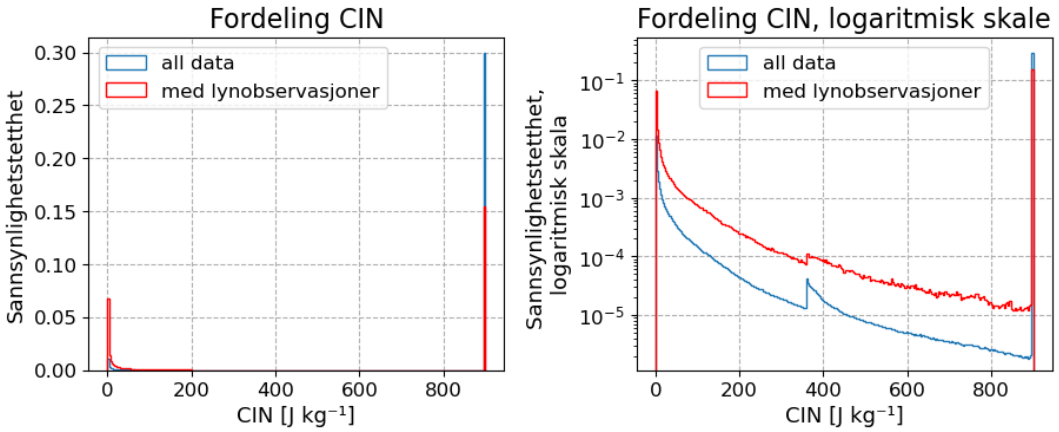
Sannsynlighetsplottene i fgurene 4.2. til 4.8 vises sannsynlighetfordelingen for alle undersøkte parameterne. Blåekurver viser sannsynlighetstetthetsfordeling over hele datasettet, mens de røde kurvene viser kun data med lynobservasjoner. Med lynobservasjoner betyr datapunkter hvor det er observert minst ett lyn. Til venstre er det normal skal, men til høyre er det semi-logaritmisk skala.

I figur 4.2 vises sannsynlighetstettheten til CAPE. De fleste verdiene for CAPE ligger rundt 0Jkg^{-1}. De høyeste verdiene når 5000 Jkg^{-1}. Det er en antydning til at det er relativt mer lyn når CAPE er høyere for pottet med normal skala, men dette kommer enda tydeligere frem i plottet med semi-logaritmisk skala. Her ser man tydelig at sannsynligheten for lyn er høyere når CAPE er høyere.



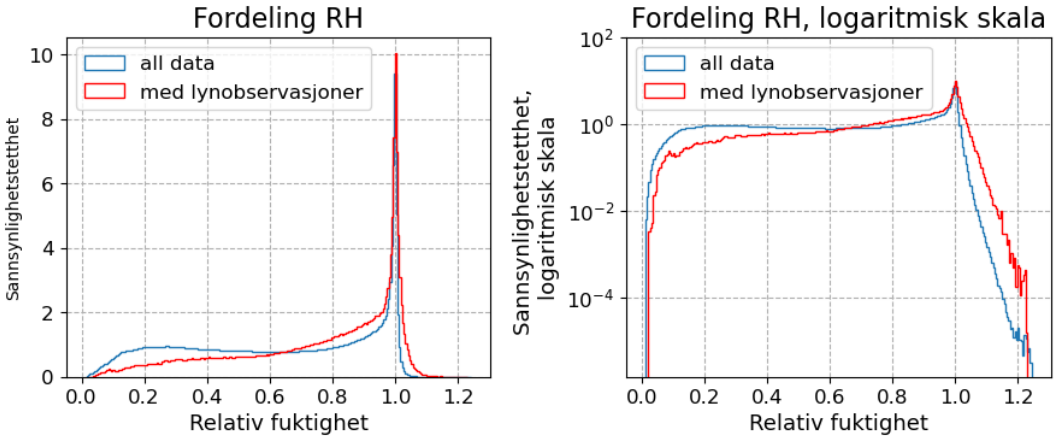
Figur 4.2: Sannsynlighetstetthetsplot av CAPE. Figuren viser fordelingen av verdier med normal skala til venstre, og semi-logaritmisk skala til høyre. Blå farge er for hele datasettet, mens rød farge er for datapunkter som inneholder lynobservasjoner. Arealet under hver av grafene er lik 1.

I figur 4.3 vises sannsynlighetstettheten til CIN. Det tydeligste er at de fleste verdiene er på 900 Jkg^{-1}, og at det er ganske mange på 0 Jkg^{-1}. På den semi-logaritmiske skalaen til høyre kommer flere nyanser frem. Her kan man se at det er relativt mer sannsynlighet for lyn når CIN er lavere. Den ellers ganske monotone kurveformen mellom minimum og maksimumsverdiene har et sprang rundt 350Jkg-1.



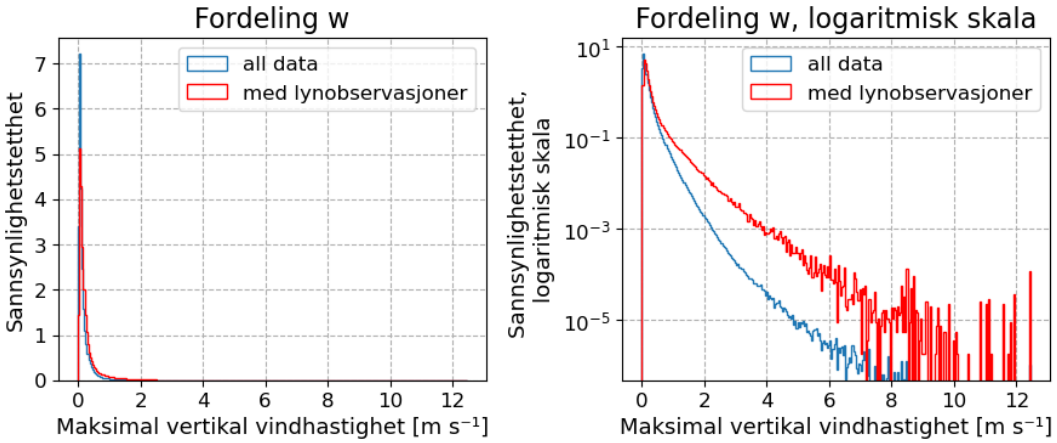
Figur 4.3: Sannsynlighetstetthetsplot av CIN. Figuren viser fordelingen av verdier med normal skala til venstre, og semi-logaritmisk skala til høyre. Blå farge er for hele datasettet, mens rød farge er for datapunkter som inneholder lynobservasjoner. Arealet under hver av grafene er lik 1.

I figur 4.4 vises sannsynlighetstettheten til RH. I begge plot kan man se at verdiene er mellom 0 og 1.2, med en topp rundt 1 og de fleste verdiene lavere enn dette. Det er ingen tydelige forskjeller på lynobservasjoner og resten av datasettet, men man kan også se en antydning til relativt høyere sannsynlighet for lyn når RH er høyere, og at det er lavere når RH er lav.



Figur 4.4: Sannsynlighetstetthetsplot av RH. Figuren viser fordelingen av verdier med normal skala til venstre, og semi-logaritmisk skala til høyre. Blå farge er for hele datasettet, mens rød farge er for datapunkter som inneholder lynobservasjoner. Arealet under hver av grafene er lik 1.

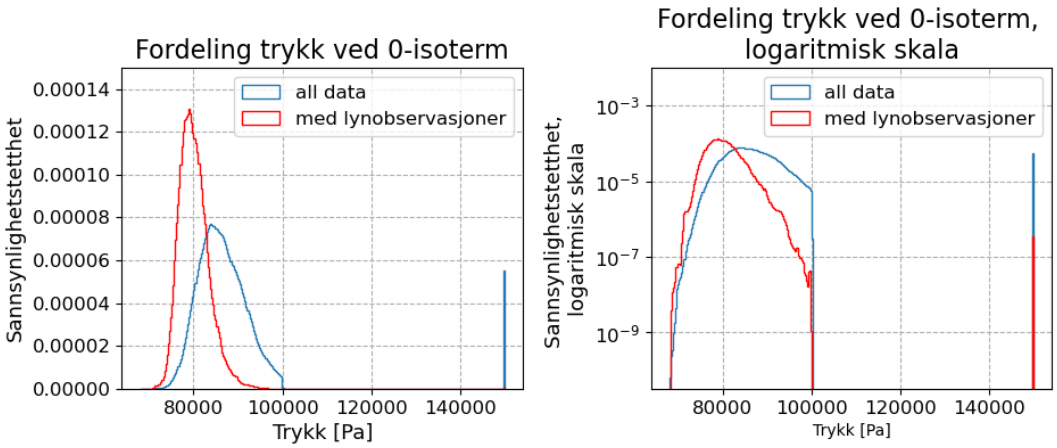
I figur 4.5 vises sannsynlighetstettheten til w. Verdiene her er mellom 0 og 12 ms^{-1}, de fleste rundt 0. Spesielt i plottet til høyre, med semi-logaritmisk skala, kan man se at høyere vindhastighet har relativt sett høyere sannsynlighet enn lyn.



Figur 4.5: Sannsynlighetstetthetsplot av w. Figuren viser fordelingen av verdier med normal skala til venstre, og semi-logaritmisk skala til høyre. Blå farge er for hele datasettet, mens rød farge er for datapunkter som inneholder lynobservasjoner. Arealet under hver av grafene er lik 1.

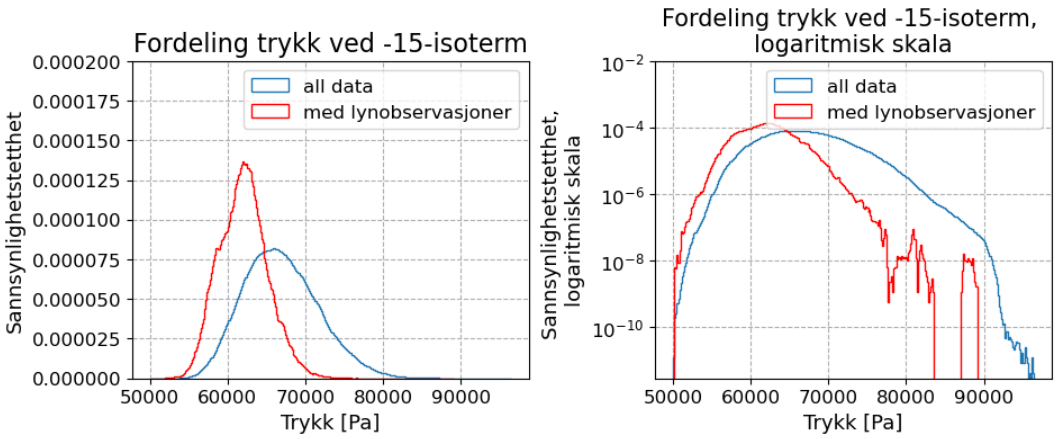
I figur 4.6 vises sannsynlighetstettheten til høyden til 0-isotermen. Både hele datasettet og tilfellene med lyn ser ganske normalfordelt ut. Sannsynlighetstettheten for lynobservasjoner er forskjøvet noe til venstre, mot lavere trykk. Spesielt i plottet til høyre med semi-logaritmisk skala ser man tydelig at verdiene blir kuttet av ved 100000Pa, og at det er flere verdier ved 150000Pa.

I figur 4.7 vises sannsynlighetstettheten til høyden til -15-isotermen. Her er også verdiene normalfordelt, men sannsynlighetstettheten til lynobservasjoner er forskjøvet til venstre



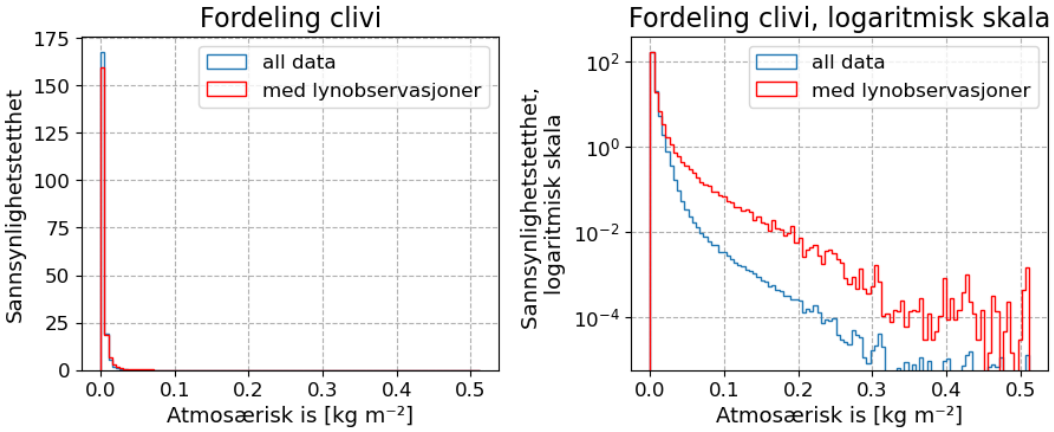
Figur 4.6: Sannsynlighetstetthetsplot av høyden 0-isotermen i Pa. Figuren viser fordelingen av verdier med normal skala til venstre, og semi-logaritmisk skala til høyre. Blå farge er for hele datasettet, mens rød farge er for datapunkter som inneholder lynobservasjoner. Arealet under hver av grafene er lik 1.

mot lavere trykk. Her er verdiene mellom 50000 og 100000Pa, uten å bli avkuttet eller med verdier utenfor normalfordelingen, men det mangler lynobservasjoner rundt 85000Pa.



Figur 4.7: Sannsynlighetstetthetsplot av høyden til -15-isotermen i Pa. Figuren viser fordelingen av verdier med normal skala til venstre, og semi-logaritmisk skala til høyre. Blå farge er for hele datasettet, mens rød farge er for datapunkter som inneholder lynobservasjoner. Arealet under hver av grafene er lik 1.

I figur 4.8 vises sannsynlighetstettheten til clivi. Verdiene er mellom 0 og 0.5 kgm^{-2}, og det er flest verdier rundt 0. I plottet til venstre kan man se at det er større sannsynlighetstetthet for lyn når clivi er høy.



Figur 4.8: Sannsynlighetstetthetsplot av clivi. Figuren viser fordelingen av verdier med normal skala til venstre, og semi-logaritmisk skala til høyre. Blå farge er for hele datasettet, mens rød farge er for datapunkter som inneholder lynobservasjoner. Arealet under hver av grafene er lik 1.

4.1.3 Multipliserte parametere

I figur 4.9 vises tetthetsfordelingen for høyden til 0-isotermen multiplisert med de andre parameterne. Parameterne er først skalert til mellom 0 og 1. Se vedlegg B for figurer av alle parametere multipliserte med hverandre.

I a) ser man 0-isotermen x CIN. Datapunktene er skiftet litt til venstre for hele datasettet, og sannsynlighetstettheten er høyere på de lavere verdiene. Det er ikke en like tydelig forskyvning som for kun 0-isotermen, men tydeligere enn for CIN. Ved 1 er det separate datapunkter. Disse kommer av at både CIN og 0-isotermen opprinnelig har NaN-verdier som har blitt substituert med henholdsvis 900Jkg^{-1} og 150000 Pa.

I b) vises 0-isotermen x CAPE. Her er sannsynlighetstettheten til lynobservasjoner lavest rundt 0, men ligger jevnt litt over det totale datasettet ellers.

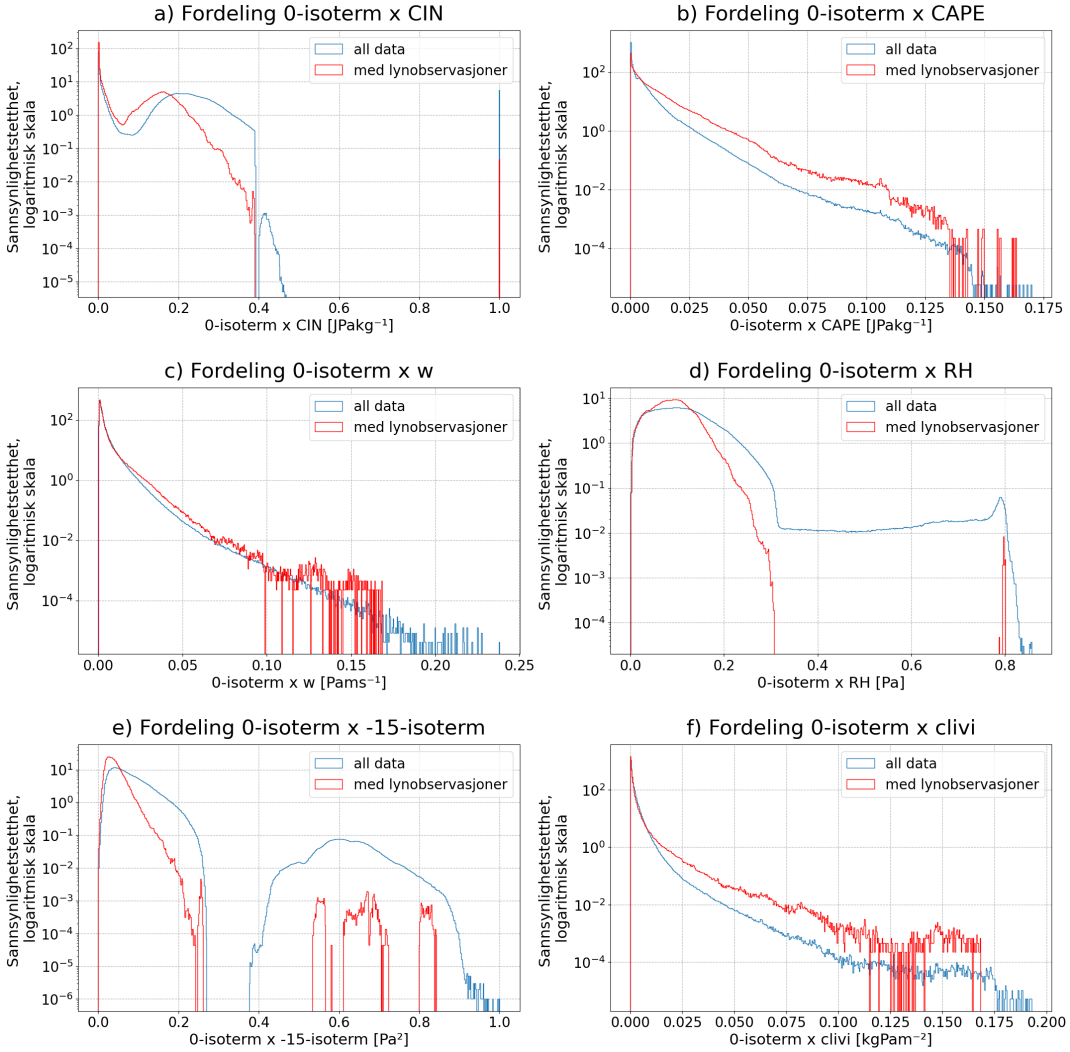
I c) vises 0-isotermen x w. Her er er sannsynlighetstettheten for lynobservasjoner ganske lik resten av datasettet, men litt høyere mellom 0,1 og 0,17. Ved verdier større 0,17 blir ingen lynobservasjoner registrert

I d) vises 0-isotermen x RH. Sannsynlighetstettheten til lynobservasjonene for denne kombinasjonen minner veldig om 0-isotermen, med ganske normalfordelt data ved lavere verdier. Toppen av fordelingen ligger rundt 0.1. Det er en ny topp ved de høyeste verdiene, rundt 0.8. Mellom 0.3 og 0.8 er det ingen lynobservasjoner, mens sannsynlighetsfordelinger til hele datasettet ligger jevnt på 0,01.

I e) vises 0-isoterm x -15-isoterm. Her er dataene samlet i to fordelinger. Den ene rundt 0.1 og den andre rundt 0.6. Sannsynlighetstettheten til lynobservasjonene ligger stort sett lavere enn resten av datasettet, med untak for de helt laveste verdiene hvor de har

høyere sannsynlighetstetthet.

I f) vises 0-isotermen x clivi. Her ligger sannsynlighetstettheten jevnt litt høyere enn for det totale datasettet.

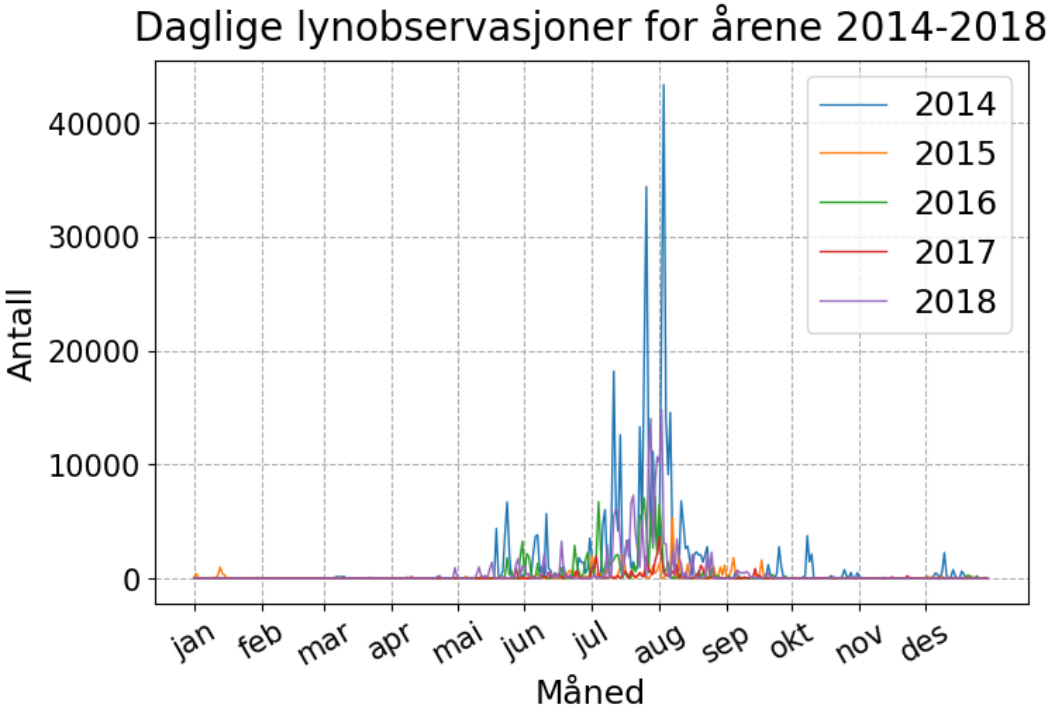


Figur 4.9: Sannsynlighetstetthetsplot av fordelingen av alle parametere multiplisert med 0-isotermen i en semi-logaritmisk skala. a) viser 0-isotermen×CIN, b) viser 0-isotermen×CAPE, c) viser 0-isotermen×w. d) viser 0-isotermen×RH, e) viser 0-isotermen×-15-isotermen, f) viser 0-isotermen×clivi. Blå farge er for hele datasettet, mens rød farge er for datapunkter som inneholder lynobservasjoner. Arealet under hver av grafene er lik 1.

4.1.4 Lyndata

De observerte lyndataene ble undersøkt. Dette er lynobservasjonene ved de synoptiske tidspunktene og har en tidsoppløsning på tre timer.

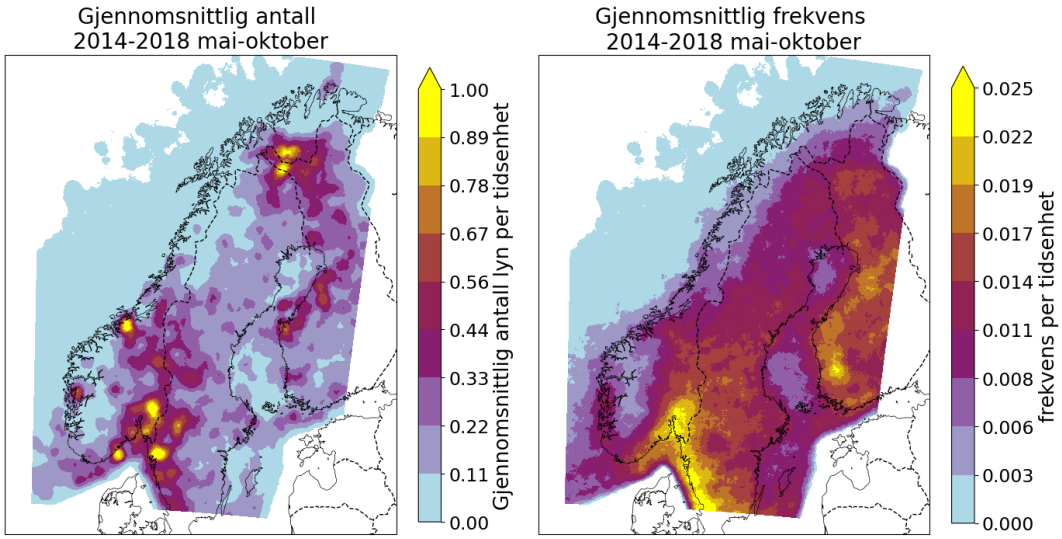
Det totale antallet lyn daglig for hele utstrekningen til kartutsnittet for de fem årene 2014-2018 vises i figur 4.10. For samtlige år er det mest aktivitet mellom mai og september, spesielt mye i juli og august, men det er store variasjoner mellom hvert år. 2014 utpeker seg som et år med ekstra mye aktivitet, mens 2017 har minst. De andre årene varierer. 2014 og 2015 ser også ut til å ha mer aktivitet enn de andre årene på vinteren, med noe aktivitet i desember og januar.



Figur 4.10: Sum per dag av antall lynobservasjoner ved de synoptiske tidspunktene. Summen er over hele utstrekningen til kartutsnittet og dekker årene 2014-2018.

I figur 4.11 vises gjennomsnittlig antall lyn for 2014-2018, mai-oktober til venstre, og gjennomsnittlig frekvens per tidsenhet time til høyre. Figuren til venstre tar antall lyn med i beregningen, mens figuren til høyre kun ser på om det er registrert en eller flere lynhendelser ved gitt tidspunkt. Figuren til venstre viser mest aktivitet rundt østlandet, Nordmøre/Trøndelag og sør på Finnmarksvidda/nord i Finnland/Sverige med opptil 1 lyn i gjennomsnitt per 3. time. Det er lavest gjennomsnitt over nordsjøen, på Sunnmøre og i Nordland, kysten av Troms, kysten og store deler av Finnmark og sør på/for Hardangervidda. Mellom grensa til Norge og Sverige er det middels høyt gjennomsnitt på rundt 0.5, og rundt Bergen er gjennomsnittet over 0.5, men ikke helt opp i 1.

Fordelingen endrer seg når man kun betrakter lynobservasjoner som en boolsk variabel. Kartet til høyre i figur 4.11, hvor antall lyn ikke blir tatt med i beregningen, men kun om det er lyn eller ikke, vise et annerledes mønster. Det er høyest frekvens på Østlandet, og kysten på sør-østlandet og sør-vest i Sverige. Det er også en ganske høy frekvens sør i Finnland. Frekvensen er lavest over nordsjøen, og øker gradvis inn mot kysten av Sør-Norge og innover i landet. I Nord-Norge er det langs kysten og ganske langt inn på fastlandet. I Trøndelag, på Vestlandet og på Sørlandet er frekvensen noe høyere, men ikke så høy som på Østlandet. I fjellstrøkene vest i Sør-Norge og Sunnmøre er frekvensen ganske lav.



Figur 4.11: Til venstre vises gjennomsnittlig antall lyn ved de synoptiske tidspunktene, og til høyre vises frekvens av lynhendelser ved de synoptiske tidspunktene. Begge plottene er fra mai-oktober for 2014-2018 og er fra observasjonsdata.

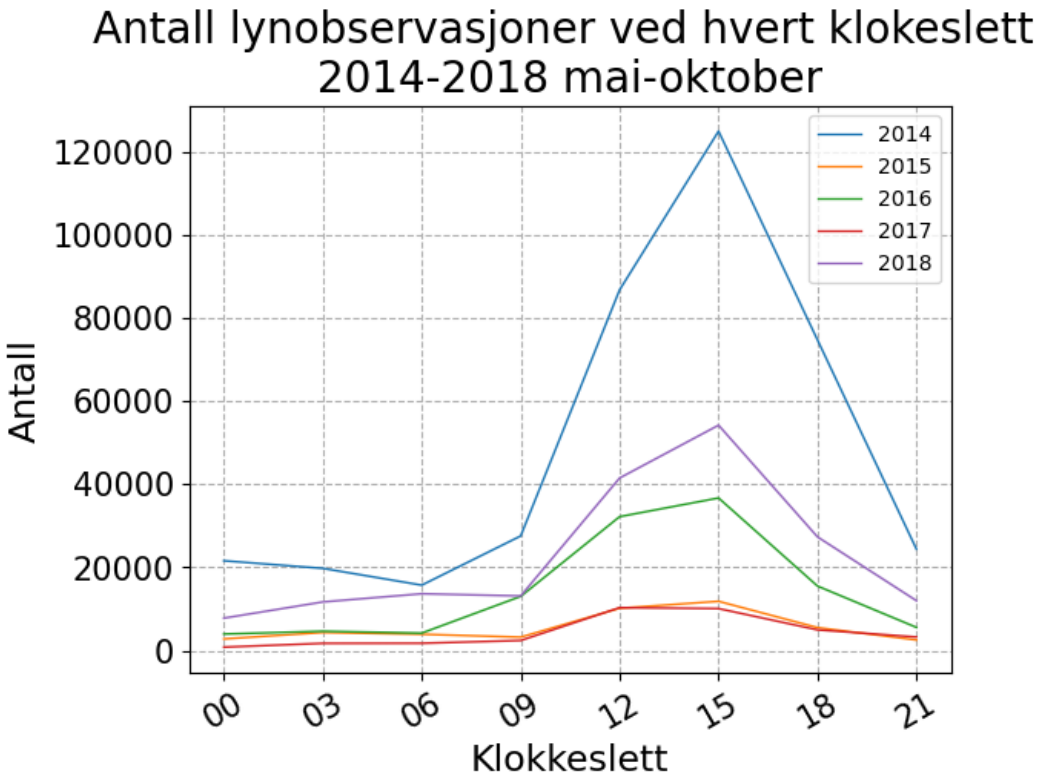
I figur 4.12 vises ved hvilke tidspunkt det har blitt registrert lyn. Det er en tydelig topp på ettermiddagen for alle år, mens det er mye mindre aktivitet på natta.

4.2 Preprosessering

Noen av stegene underveis i preprosesseringa blir presentert her. Mange av prosessene ble kjørt samtidig, så det finnes ikke info om alle mellomstegene, men det som er tilgjengelige blir visualisert.

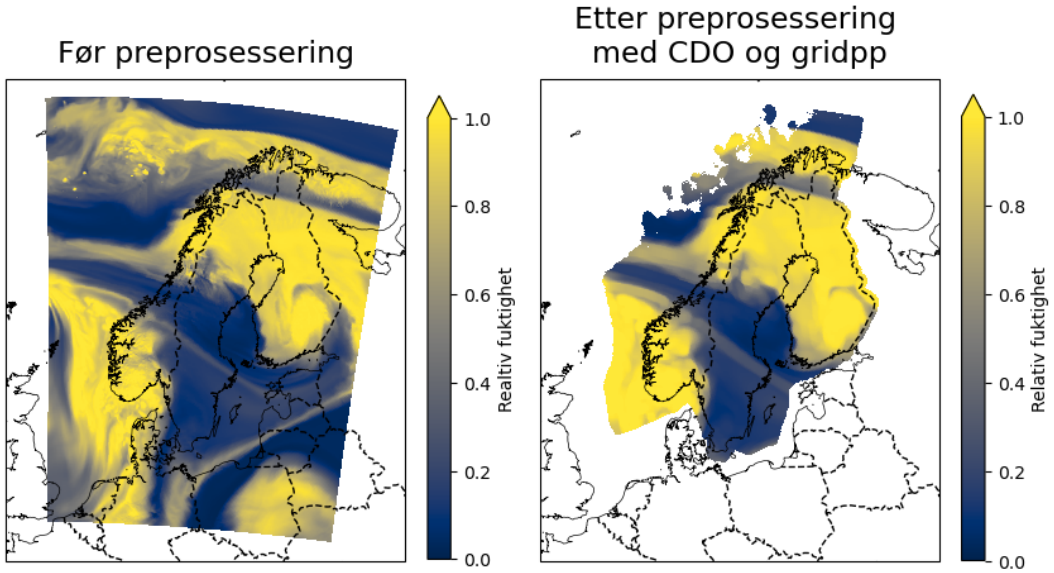
4.2.1 CDO og gridpp

Del 1 av preprosesseringa, som gjøres med CDO og gridpp, er beskrevet i delkapittel 3.2.2. I figur 4.13 vises et eksempel med RH av hvordan dataene går fra å dekke et stort område og være ubehandlet, til å være redusert i utstrekning og med glattede



Figur 4.12: Summen av antall lynobservasjoner ved de synoptiske tidspunktene. Summen er over hele utstrekningen til kartutsnittet og dekker årene 2014-2018.

verdier. RH beregnes først fra spesifikk fuktighet ettersom det ikke er et parameter i klimadataene.



Figur 4.13: Visuelt eksempel på del 1 av preprosessering av parametere som gjøres med CDO og gridpp. Til venstre vises RH, beregnet ut i fra spesifikk fuktighet, trykk og temperatur, før resten av preprosesseringa er gjort. Til høyre vises RH etter å ha blitt glattet ut med gridpp, og begrenset til et utsnitt av området som dekkes av lynsensorene. Eksempelet er fra 01.01.2014 kl 03:00.

4.2.2 Redusering av datasettet

Videre ble datasettet redusert i størrelse. I tabell 4.1 viser en oversikt over filstørrelsen på hvert år underveis i preprosesseringa for treningsdataene. Siden ikke alle steg underveis i preprosesseringa har en egen fil er det kun deler av prosessen hvor filstørrelsen er kjent. Filstørrelsen er gitt i antall bytes som trengs for å lagre dataene.

I kolonne A er filstørrelsen etter at datasettet har gått gjennom del 1 av preprosesseringa med CDO og gridpp. Data forligger i netcdf-filer som består av parameterne CAPE, CIN, RH, w, 0-isoterm, -15-isoterm, clivi, lynobservasjoner i hver gridcelle, lynobservasjoner innenfor 25km av hver gridcelle, land/vann, lengdegrader og breddegrader. Dataene omfatter hele året og verdiene er maskert til området med lynobservasjoner. Dette er den største størrelsen datasettet her.

Kolonne B viser filstørrelsen etter reduseringen. Datasettet foreligger nå som Dataframe og inneholder kun månedene mai-oktober. Videre er en del av området rundt Norge fjernet. Datasettet består av de samme parameterne som før, inkludert lengde- og breddegrad, dag i året og tidspunkt. Det er disse datasettene som brukes til å predikere lynobservasjoner og å visualisere dem i kart.

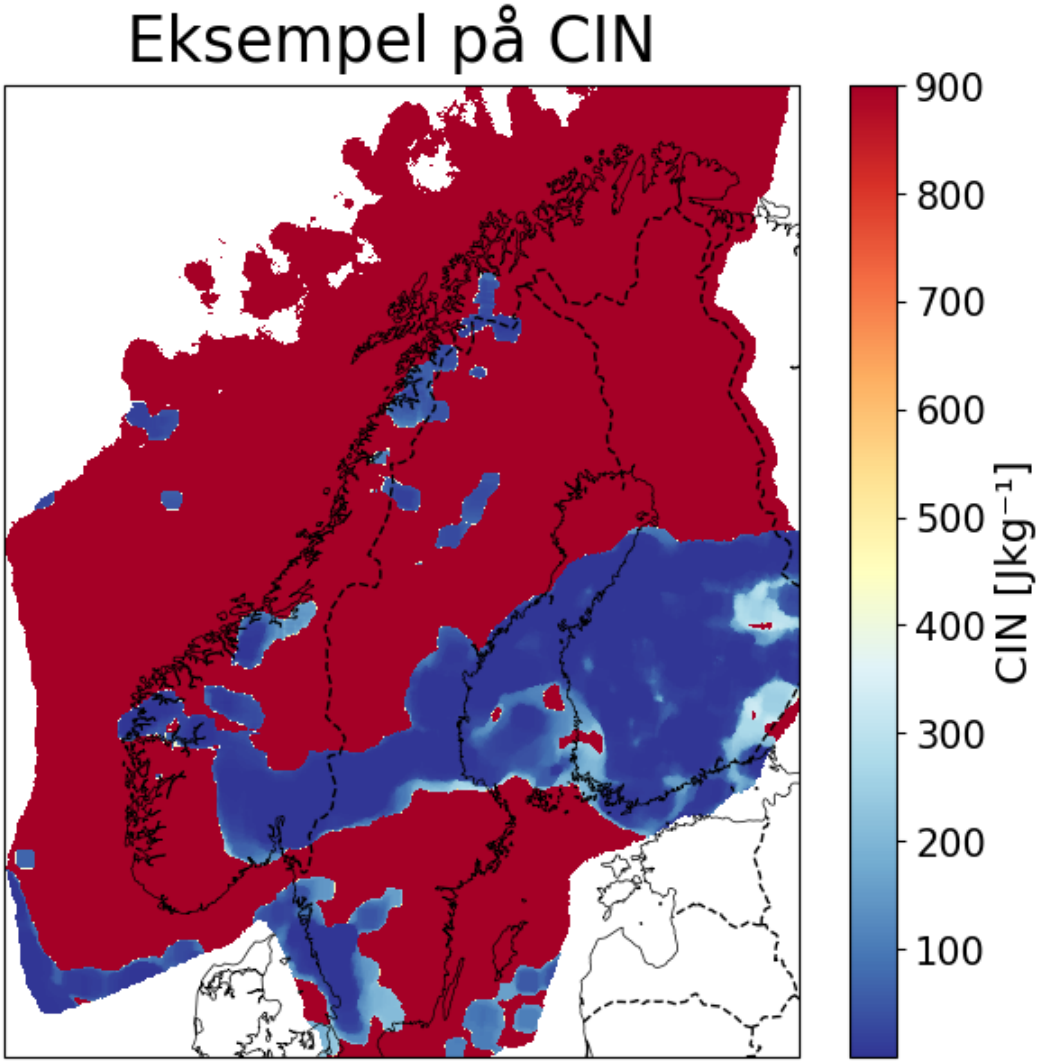
Kolonne C viser filstørrelsen etter en ytterligere redusering. Indeksen i disse filene inneholder nå ikke lenger info om lengde- og breddegrad, men har en numerert ID. Siden datasettet opprinnelig er griddet er det flere rader som inneholder NaN-verdier fordi de er utenfor området med lynobservasjoner. Disse blir fjernet her. 75% av CIN verdier er fjernet for hver måned. I figur 4.14 vises et eksempel på hvor mye av datasettet som kan inneholde CIN = 900 Jkg^{-1}. Datasettet ved punkt C er utgangspunktet for trening og utprøving av modeller.

Fra A til C reduserer filstørrelsene til omtrent 1/10. Det er også preprosessert og klart til å brukes i maskinlæring.

Tabell 4.1: Tabell som viser hvordan størrelsen på treningsdataene endret seg gjennom reduksjon av filstørrelsen. I kolonne A vises antall bytes som trengs for å lagre datasettet etter del 1 av preprosesseringa, i B vises antall underveis i nedskaleringa og kolonne C viser antall bytes som trengs for å lagre datasettet etter at nedskaleringen er ferdig.

Årstall	A	B	C
2014	72.9G	33.5G	7.9G
2015	72.9G	33.5G	6.7G
2016	73.1G	33.5G	7.0G
2017	72.9G	33.5G	6.9G
2018	72.9G	33.5G	7.5G

Siden det lyn er et sjelden fenomen, er antall datapunkt uten lyn mange størrelsesorden



Figur 4.14: Eksempel på hvor mye av CIN som har verdien 900Jkg^{-1}. Rød farge er CIN=900Jkg^{-1}, mens blå er lavere verdier. Tidspunkt er 01.08.2014 kl. 12:00

større enn datapunkt med lyn. Dette gjør datasettet veldig ubalansert. I tabell 4.2 vises prosent med datapunkter som inneholder lynobservasjoner. A, B og C refererer til samme steg i prosessen som for tabell 4.1. Fra A til C øker den relative andelen med lyn med rundt 10 ganger så mye.

Tabell 4.2: Prosent lyn i treningsdataene ved forskjellige tidspunkter i prosessen for å redusere størrelsen på datasettet. Kolonne A viser prosent lyn i datasettet etter del 1 av preprosesseringa, som dekker januar-desember. Kolonne B viser prosent underveis, når blant annet perioden er begrenset til mai-oktober og C viser prosenten når reduksjonen er ferdig.

Årstall	A [%]	B [%]	C [%]
2014	0.561	1.828	6.203
2015	0.130	0.310	1.233
2016	0.187	0.606	2.319
2017	0.091	0.272	1.058
2018	0.150	0.519	1.875

4.3 Maskinlæringsmodeller

Å utvikle maskinlæringsmodeller produserer mange resultater. I dette delkapittelet presenteres de mest sentrale resultatene relatert til maskinlæringsmodellene.

4.3.1 Tidlig utprøving av modeller

I tabell 4.3 vises en oversikt over de forskjellige modellene som ble utprøvd for å finne hvilkenkombinasjon av parametre som ga best resultater. Alle parametrene er skalert med min-max mellom 0 og 1, og året som er brukt til trening og testing er 2015. «logreg» står for logistisk regresjon, «rfc» står for RandomForestClassifier, og tallene representerer hvilken kombinasjon av parametre som er brukt. To lineære logaritmiske modeller ble utprøvd. Én med parameterne CAPE, CIN, RH og w, og betegnes som modell-0, og én med alle tilgjengelige parametre. Det ble utprøvd 15 forskjellige kombinasjoner av rfc-modeller. De ble vurdert etter ytelsesmålene forklart i delkapittel 3.3.2. Opprinnelig var f1-score tenkt som eneste ytelsesmål, men fra og med modell 4 ble multipliserte parametre utprøvd. Dette førte til at den generelle sannsynligheten for lyn ble signifikant lavere. Siden f1-score bruker grenseverdi 0.5 med mindre annet er spesifisert, gjorde dette at f1-scoren til modeller med multipliserte parametre ble omtrent en størrelsesorden laverer enn for modeller med ikke-kombinerte parametre, selv om ROC-AUC scoren forble ganske lik de modellene som ikke hadde multipliserte parametre.

Av de modellene som ikke inneholdt multipliserte parametre var logreg_0 lavest både på f1-score og ROC-AUC, henholdsvis 0.1746 og 0.710(trening) og 0.711(test). Det er en logistisk regresjon med parametrene CAPE, CIN, RH og w. Den modellen med høyest f1-score og ROC-AUC var rfc_3, med henholdsvis 0.603 og 0.924(trening) og 0.922(test). Dette er en Tilfeldig skog klassifiseringssystem med alle tilgjengelige parametre; CAPE, CIN, RH, w, 0-isoterm, -15-isoterm, clivi, landmaske og tidspunkt. Resterende modeller

viser at rfc generelt gjør det bedre enn logreg, og at flere parametre gir bedre resultater.

Av de modellene med multipliserte parametre, fikk rfc_5 det dårligste resultatet med en f1-score på 0.0377 og ROC-AUC på 0.894(trening) og 0.891(test). Denne modellen inneholdt flest multipliserte parametre, CIN x 0-isoterm, 0-isoterm x -15-isoterm, w x clivi og RH x 0-isoterm. CAPE var eneste kontinuerlige parameter som ikke var multiplisert, i tillegg til at de kategoriske parametrene landmaske og tidspunkt ble brukt. rfc_4 gjorde det best av de med multipliserte parametre, med en f1-score på 0.0754 og ROC-AUC på 0.916(trening) og 0.914(test).

Den videre analysen med det komplette dataset ble gjort på et representativ utvalg av modeller: logreg_0, rfc_3, rfc_4, rfc_10, rfc_11 og rfc_15.

4.3.2 Videre utvikling av enkelte modeller

De utvalgte modellene ble utviklet videre. De ble blant annet trent og testet på hele datasettet, og flere ytelsesmål ble undersøkt.

I figur 4.4 vises ROC-AUC målene for de modellene som ble undersøkt videre. Her er treningen gjort for et utdrag av hele datasettet, der 30% har blitt utelatt fra treningen og brukt til å teste. Dette er for å sjekke om modellene overtilpasser. Det er lite forskjeller mellom ROC-AUC scorene for alle modellene i 4.4.

I tabell 4.5 vises forskjellige ytelsesmål for modeller som er trent på all data. Logreg_0 har lavest verdi på ROC-AUC, mer en 0,1 (eller nesten 0.2) lavere enn alle andre modellene. Rfc_3 får høyest score med 0.891. Se vedlegg C for figur av ROC-kurven. Grenseverdiene er de verdiene som er funnet ved å bruke kvantiler, forklart i delkapittel 3.3.2. De modellene som kun har de originale parameterne har grenseverdier på rundt 0.8, mens de som også inkluderer parametere som er multiplisert sammen har grenseverdier rundt 0.3. Disse verdiene sier ikke noe om hvor presise modellen er, men mer om hvor høy den generelle sannsynligheten for lyn er.

Rangering av år land og vann viser hvor mange år de forskjellige modellene klarte å rangere etter observertlynfrekvens. Ingen av modellene klarte å rangere mer enn tre av årene riktig, og flere gjorde flere feil for celler over vann.

F1-scoren er beregnet ut i fra TP, TN, FP og FN for de sammsynlighetsberegningenen modellene har gjort som er over/under de respektive grenseverdiene. Rfc_10 har den høyeste f1-scoren, med 0.155. Rfc_11 har den laveste med 0.0961. De modellene med kun originale parametere gjør det generelt bedre, mens de med multipliserte parametere får lavere f1-score.

Tabell 4.3: Oversikt over forskjellige modeller og ytelsesmål. Modellene er både logaritmisk regresjon og rfc, og har forskjellige kombinasjoner av parametere. Hvor bra de presterer blir vurdert ut i fra ytelsesmålene f1-score, ROC-AUC trening og ROC-AUC test. Jo høyere disse verdiene, jo bedre presterer modellene.

Modell	parametre	f1-score	ROC-AUC trening	ROC-AUC test
logreg_0	CAPE, CIN, w, RH	0.1746	0.710	0.711
logreg_3	CAPE, CIN, RH, w, 0-isoterm, -15-isoterm, clivi, landmaske, tid	0.237	0.828	0.829
rfc_0	CAPE, CIN, RH, w	0.528	0.795	0.792
rfc_1	CAPE, CIN, RH, w, 0-isoterm, -15-isoterm, clivi, landmaske	0.567	0.901	0.900
rfc_2	CAPE, CIN, RH, w, 0-isoterm, -15-isoterm, clivi, tid	0.599	0.920	0.919
rfc_3	CAPE, CIN, RH, w, 0-isoterm, -15-isoterm, clivi, landmaske, tid	0.603	0.924	0.922
rfc_4	CAPE, CIN, RH, w, 0-isoterm, -15-isoterm, clivi, landmaske, tid, CIN*0-isoterm	0.0754	0.916	0.914
rfc_5	CAPE, landmaske, tid, CIN*0-isoterm, 0-isoterm*-15-isoterm, w*clivi, RH*0-isoterm	0.0377	0.894	0.891
rfc_6	CAPE, RH, w, tid, CIN*0-isoterm, 0-isoterm*-15-isoterm	0.0458	0.899	0.895
rfc_7	CAPE, tid, RH, w, landmaske, CIN*0-isoterm, 0-isoterm*-15-isoterm	0.0500	0.901	0.899
rfc_8	CAPE, w, RH, 0-isoterm	0.572	0.895	0.893
rfc_9	CAPE, w, RH, 0-isoterm, tid, landmaske	0.574	0.900	0.899
rfc_10	CAPE, CIN, w, RH, 0-isoterm, -15-isoterm, landmaske, tid	0.59	0.913	0.912
rfc_11	CAPE, w, RH, 0-isoterm, landmaske, tid, CIN*0isoterm	0.0457	0.895	0.893
rfc_12	CAPE, CIN, w, RH, 0-isoterm, -15-isoterm, tid	0.587	0.910	0.909
rfc_13	CAPE, w, RH, 0-isoterm, tid, CIN*0-isoterm	0.0486	0.899	0.887
rfc_14	CAPE, CIN, w, RH, 0-isoterm, -15-isoterm, tid, CIN*0-isoterm	0.0663	0.907	0.908
rfc_15	CAPE, CIN, w, RH, 0-isoterm, -15-isoterm, tid, landmaske, CIN*0-isoterm	0.0659	0.912	0.907

Tabell 4.4: ROC-AUC score for modeller trent på et utdrag av data fra årene 2014-2018. ROC-AUC Trening viser scoren for treningsdataene, mens test viser for testdata som ikke har vært en del av treningsprosessen.

	ROC-AUC	
Modell	Trening	Test
logreg_0	0.671	0.671
rfc_3	0.854	0.854
rfc_4	0.848	0.847
rfc_10	0.851	0.850
rfc_11	0.843	0.842
rfc_15	0.849	0.845

Tabell 4.5: Forskjellige ytelsesverdier og grenseverdier for modeller trent på all data fra mai-oktober 2014-2018. Jo høyere ROC-AUC trening og f1-score er, jo bedre presterer modellen. Grenseverdiene beskriver hvilken verdi man må sette for prediksjon for at modellen skal ha en like stor gjennomsnittlig frekvens som observerte lyndata. Rangering viser hvor bra de forskjellige modellene rangerer 2014-2018 etter lynaktivtet. Skalaen er fra 0-5, der 0 er ingen år riktig rangert, mens 5 er alle år riktig rangert.

Modell	ROC-AUC trening	grenseverdi land	grenseverdi vann	rangering år land	rangering år vann	f1-score trening
logreg_0	0.762	0.824	0.856	3	3	0.130
rfc_3	0.891	0.865	0.793	3	3	0.150
rfc_4	0.857	0.364	0.312	3	1	0.107
rfc_10	0.889	0.855	0.775	3	3	0.155
rfc_11	0.862	0.362	0.295	3	1	0.0961
rfc_15	0.863	0.378	0.306	3	2	0.101

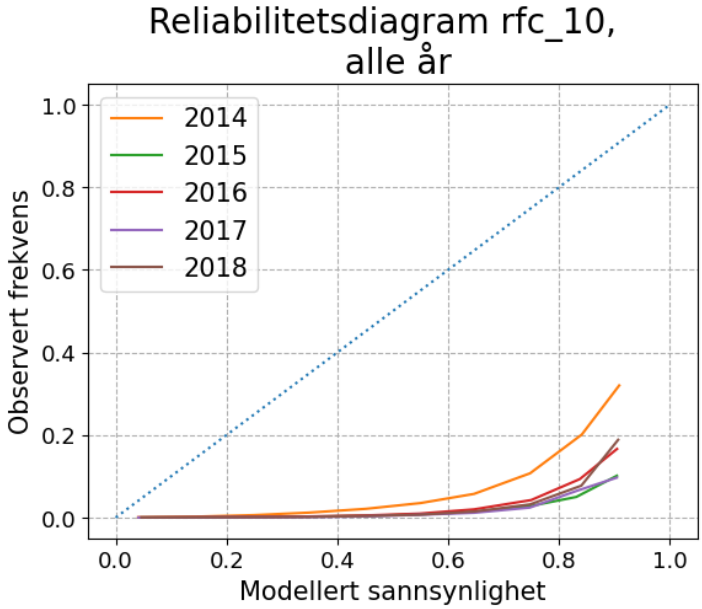
4.3.3 Reliabilitetsdiagram

I figurene 4.15, 4.16 og 4.17 vises reliabilitetsdiagrammene til henholdsvis rfc_10, rfc_11 og rfc_15.

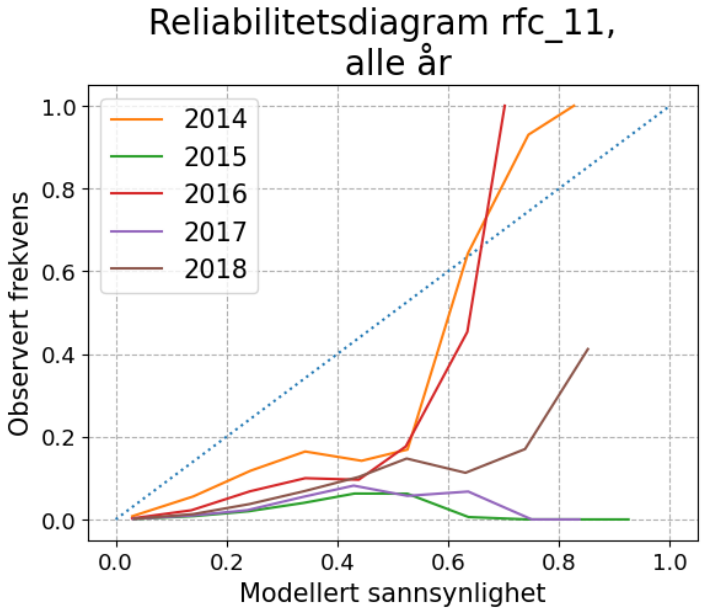
For rfc_10 i figur 4.15 er det flere modellerte høye sannsynligheter, men frekvensene til de observerte lynobservasjoner liger , og det er lite samstemthet.

For rfc_11 i figur 4.16 er det også flere høye modellerte sannsynligheter, og som stemmer bedre overens med de observerte frekvensene i forhold til de andre modellene. Grafene ligger stort sett under, men for årene 2014 og 2016, som er år med høy aktivitet, krysser grafene diagonalen.

For rfc_15 ligger alle grafene under diagonalen, men for noen av årene strekker de som opp mot 1-1-kurven.



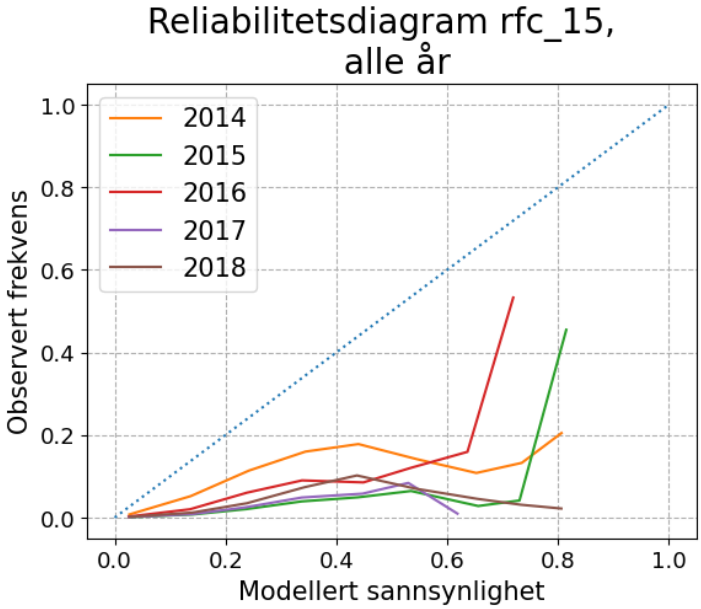
Figur 4.15: Reliabilitetsdiagram for rfc_10 som viser hvor godt predikerte sannsynligheter stemmer overens med observerte frekvenser. Ideelt skal dataene ligge langs diagonalen. Ligger de under overestimerer modellen.



Figur 4.16: Reliabilitetsdiagram for rfc_11 som viser hvor godt predikerte sannsynligheter stemmer overens med observerte frekvenser. Ideelt skal dataene ligge langs diagonalen. Ligger de under overestimerer modellen.

4.3.4 Visualisering av prediksjoner til modeller

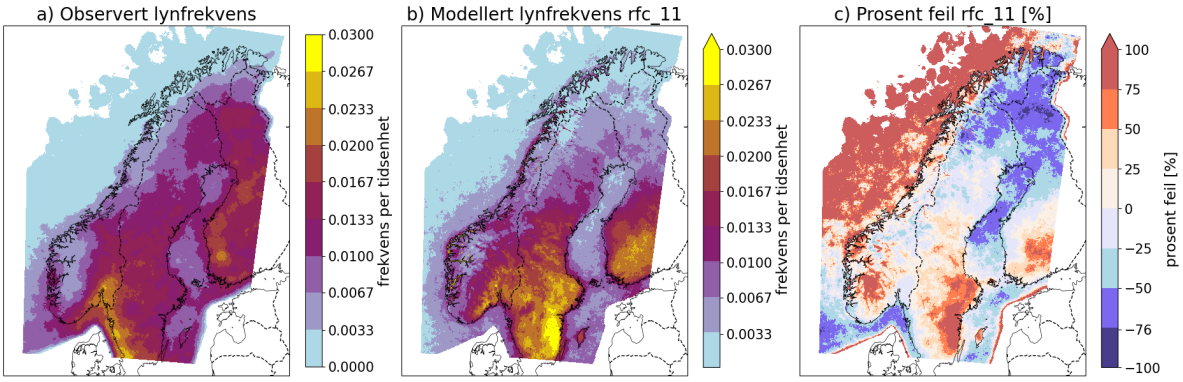
For modellene ble det laget kart for å kunne visualisere lynaktivteten. I figur 4.18 og figur 4.19 vises henholdsvis rfc_11 og rfc_15. a) plottet i begge viser gjennomsnittlig observert frekvens for hele perioden mai-oktober for alle årene 2014-2018. b) plottene viser gjennomsnittlig modellert frekvens for samme periode. I c) plottet vises hvor mange



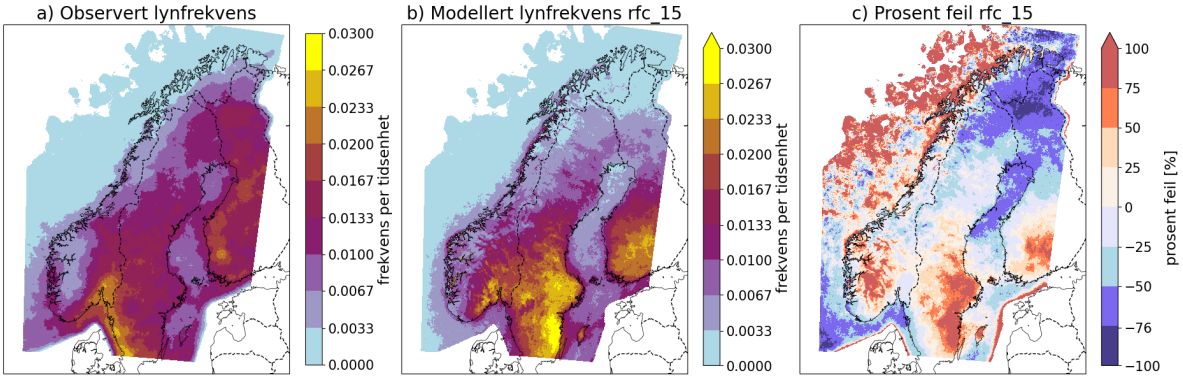
Figur 4.17: Reliabilitetsdiagram for rfc_15 som viser hvor godt predikerte sannsynligheter stemmer overens med observerte frekvenser. Ideelt skal dataene ligge langs diagonalen. Ligger de under overestimerer modellen.

prosent avvik modellene har fra det observerte.

a) er likt for begge figurene, og b) er også ganske likt. Begge har ganske lite avvik for Norge, men overestimerer fjellregionene i vest i Sør-Norge. De overestimerer også lynaktivitet i Norskehavet, spesielt rfc_11, men underestimerer i Skagerak og sør-vest for Norskekysten. For Sverige og Finnland treffer begge ganske bra, men underestimerer lengst nord, og overestimerer kraftig sør-øst i både Sverige og Finnland.



Figur 4.18: Til venstre viser observert frekvens per tidsenhet. I midten vises modellert frekvens per tidsenhet med rfc_11. Til høyre vises prosent forskjellen mellom observert og modellert frekvens.



Figur 4.19: Til venstre viser observert frekvens per tidsenhet. I midten vises modellert frekvens per tidsenhet med rfc_15. Til høyre vises prosent forskjellen mellom observert og modellert frekvens.

4.4 Modellerte endringer i lynaktivitet

For å se på endringene i lynaktivitet ble både rfc_11 og rfc_15 brukt til å modellere den historiske perioden (1986-2005), og den fremtidige perioden (2081-2100). Resultatene for modellene presenteres hver for seg.

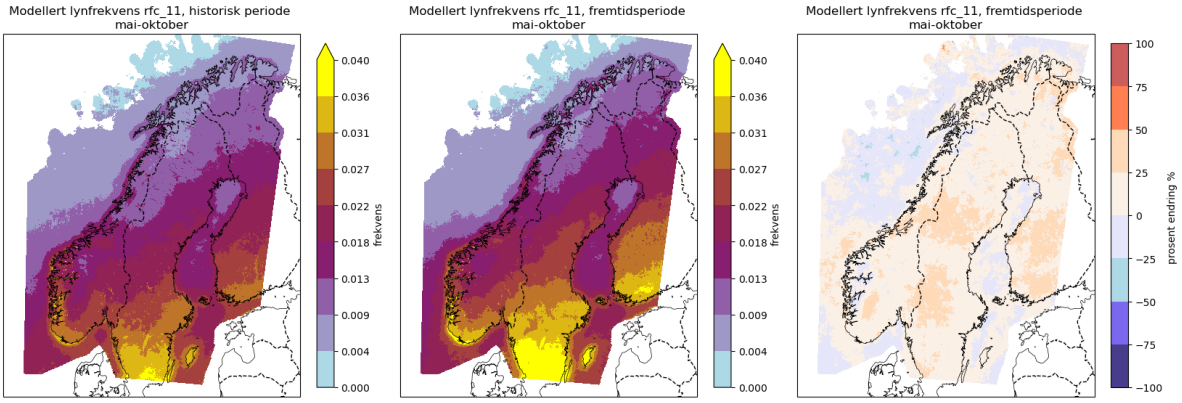
4.4.1 rfc_11

I figur 4.21 vises modellert gjennomsnittlig lynfrekvens for den historiske perioden i a), den fremtidige perioden i b), og prosent endring i c).

Både a) og b) har de samme tendensene, med høyest frekvens i sør, også gradvis økning nordover. Modellen skiller på land og vann, men det ser ut til at de kystnære områdene blir modellert til en del høyere frekvens enn de lengre ut, og det er ikke et tydelig skille mellom land og vann. I c), hvor endringen blir fremstilt, vises en generell økning på land. Stort sett mellom 0 og 25%, men noen steder opp i 50%. På Vestlandet og helt nord i Nordland er det noen steder hvor det modelleres en liten reduksjon på mellom 0 og 25%. Over vann er det både økning og reduksjon i frekvens. Vest for Sør-Norge modelleres det en økning på mellom 0 og 50 %, mens langs kysten nordover er det stort sett en reduksjon, med noe økning utenfor Troms. I Skagerak modelleres det også en reduksjon.

I modell 4.21 vises modellert prosent endring i lynfrekvensen for hver måned. Det er stor variasjon mellom månedene om det blir modellert en økning eller reduksjon i lynaktivitet.

I mai er det for det meste modellert en kraftig reduksjon i lynaktivitet, spesielt i Nordland, Troms, Nord-Sverige og -Finnland, med opptil 100% reduksjon. I Trøndelag er det både litt økning og litt reduksjon, men på Nordmøre er det et område hvor det



Figur 4.20: Resultater fra modell rfc_11. Til venstre vises modellert frekvens per tidsenhet for den historiske perioden, i midten vises modellert frekvens for den fremtidige perioden, og til høyre viser prosent endring fra historisk til fremtidig periode.

modelleres kreftig økning, opp til 100%. På Sunnmøre og Vestlandet er det både litt økning og litt redusksjon, men reduksjonen er dominerende, med noen områder som har reduksjon opp mot 100%. I Øst- og Sør-Norge er det mer moderate endringer, med de fleste verdiene mellom 25% reduksjon og 25% økning. Til havs er det store variasjoner, med både kraftige økninger og reduksjoner.

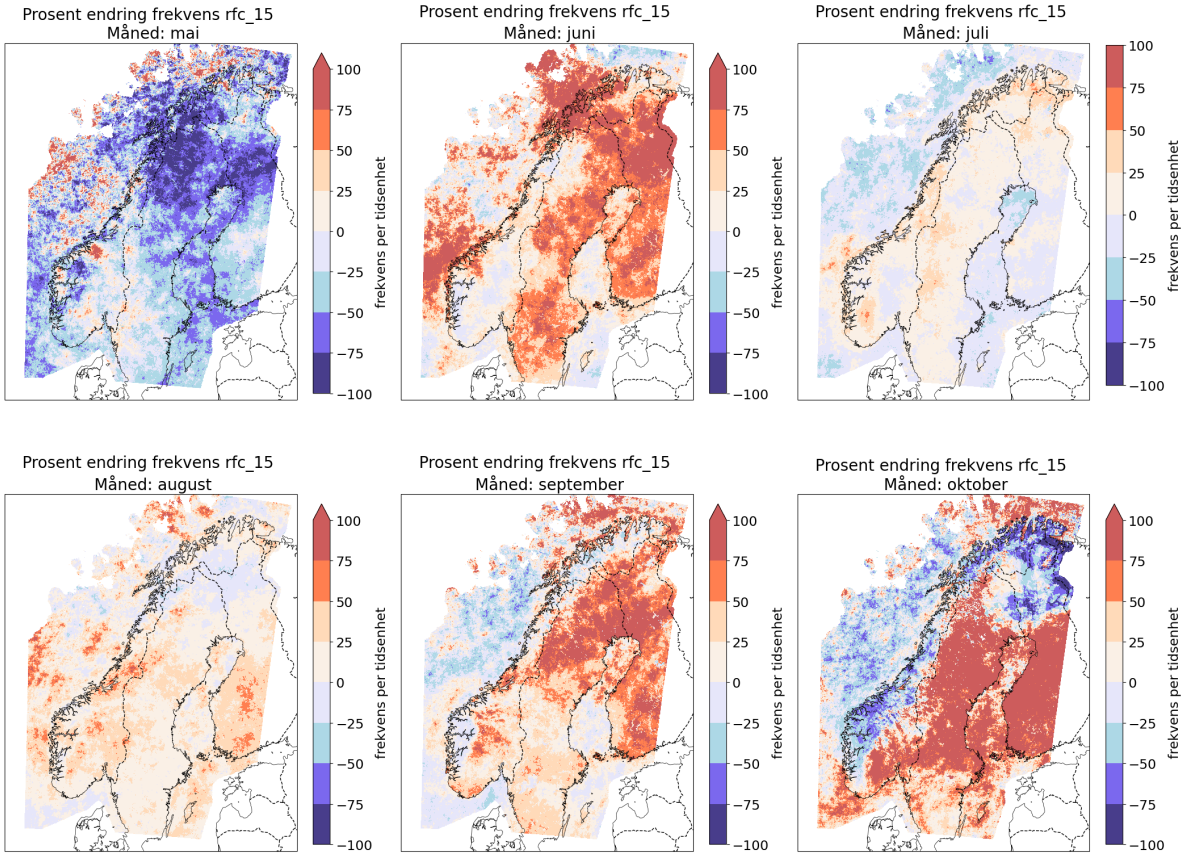
I juni modelleres det store økninger, spesiellt i Troms og over vann vest for Sør-Norge. Ellers er det moderate endringer over land, med de fleste verdiene mellom 25% reduksjon og 25% økning. Over Sunnmøre modelleres det riktignok en kraftig økning.

I juli modelleres det lite endringer i forhold til de andre månendene. De fleste verdiene er mellom 25% reduksjon og 25% økning, med noe høyere økning over Hardangervidda og i Nordland.

I august modelleres det stort sett en økning over hele fastlandsnorge. Spesielt i Trøndelag og Innlandet. Det er noe reduksjon på Finnmarksvidda og nord i Norland.

I september modelleres det kraftig økning i Innlandet, Trøndelag og langs Sørlandet. På Vestlandet, Troms og Finnmark modelleres det en reduksjon.

I oktober er det store kontraster mellom vest og øst i Norge. På Vestlandet, i Trøndelag og opplangs hele kysten nordover modelleres det en kraftig reduksjon på opptil 100%, mens øst i sør-Norge modelleres det en kraftig økning på opp mot 100%.



Figur 4.21: Resultater fra modell rfc_11. Figuren viser prosent endring i lynfrekvens per tidsenhet fra historisk til fremtidig periode. Endringene vises for månedene mai, juni, juli, august, september og oktober.

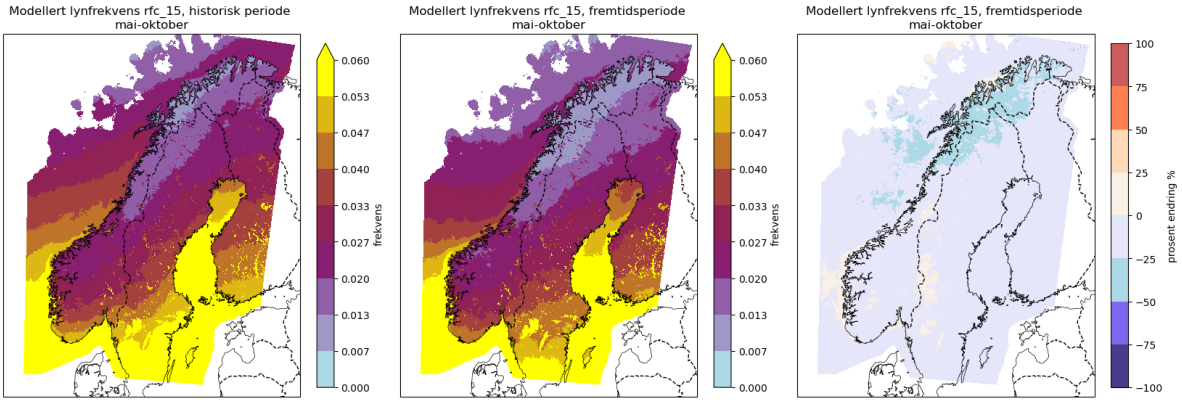
4.4.2 rfc_15

I figur 4.23 vises modellert gjennomsnittlig lynfrekvens for den historiske perioden i a), den fremtidige perioden i b), og prosent endring i c).

a) og b) viser også her høyest frekvens i sør, og en gradvis økning nordover. Her er det derimot et kraftig skille mellom land og vann, der det generelt modelleres en mye høyere frekvens over vann. I c) vises det at rfc_15 modellerer stort sett en reduksjon i lynfrekvens. Den er stort sett mellom 0 og 25%, men i Troms og Finnmark er det modellert en reduksjon på opptil 50%. Det er en modellert en liten økning langs vestkysten av Norge, på Hardangervidda og på grensa mellom Norge og Sverige, men det er små områder, og med en verdi mellom 0 og 25% økning.

I figur 4.23 vises modellert prosent endring i lynfrekvensen for hver måned.

I mai vises det en kraftig reduksjon, der store deler er oppmot 75 eller 100% reduksjon. Unntaket er øst i Sør-Norge hvor det er mer moderat reduksjon på mellom 0 og 50% reduksjon, og noen områder med økning på opp mot 50 og 75%. På Nordmøre er det også et område som har en kraftig økning på opp mot 100%.



Figur 4.22: Resultater fra modell rfc_15. Til venstre vises modellert frekvens per tidsenhet for den historiske perioden, i midten vises modellert frekvens for den fremtidige perioden, og til høyre viser prosent endring fra historisk til fremtidig periode.

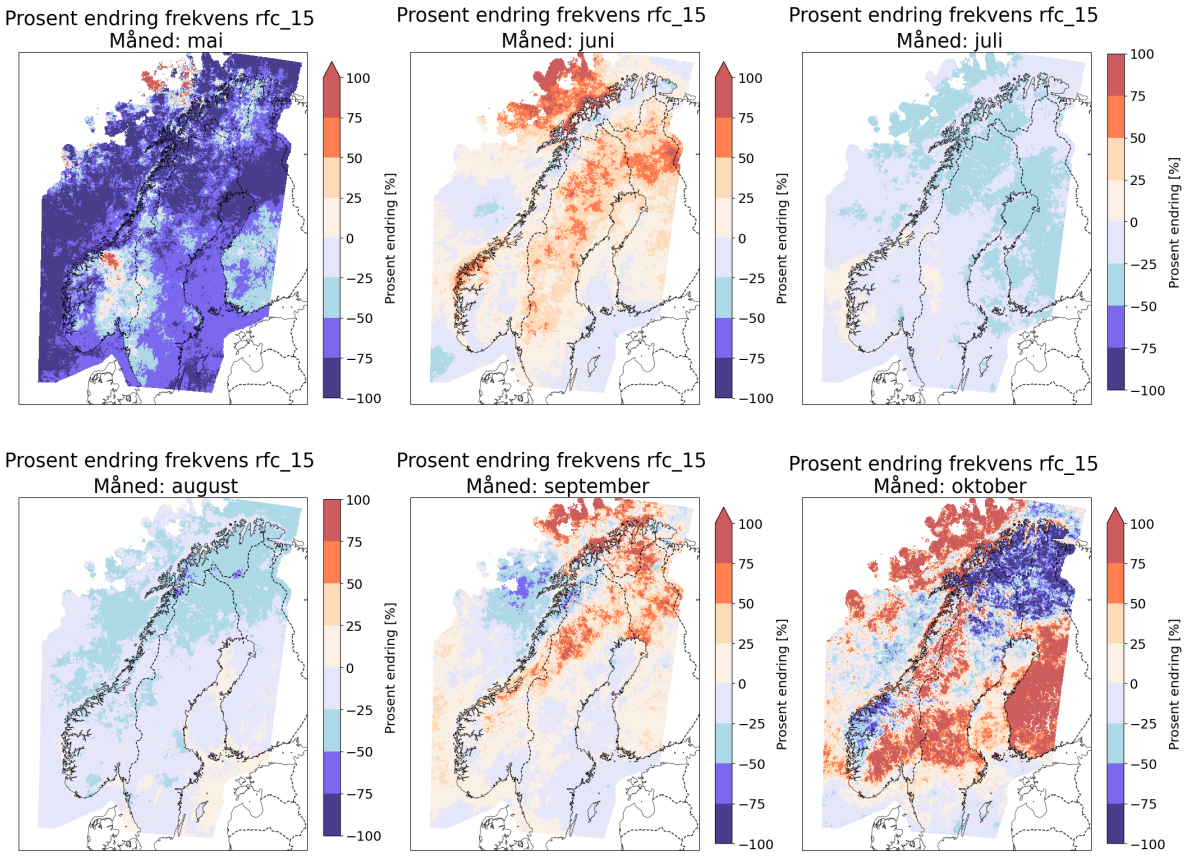
For juni vises det mer moderate endringer. Det er stort sett mellom 25% reduksjon og 25% økning over hele fastlandsnorge, med unntak av noe høyere økning langs kysten av møre hvor den er oppmot 75%.

For juli modelleres det stort sett en reduksjon på mellom 0 og 25%. I nordnorge er det noe mer reduksjon på opp mot 50%. Langs vestkysten av sørnorge modelleres det en økning på mellom 0 og 25%.

For august modelleres det nesten kun en reduksjon. Stort sett mellom 0 og 25%, men i tillegg til nordnorge er det enkelste steder i sørnorge hvor det modelleres en reduksjon oppmot 50%. Et veldig lite område på vestlandet har en modellert økning på mellom 0 og 25%.

For september modelleres det både økning og reduksjon. Det modelleres en økning for sør, og vestlandet, Innlandet, Trøndelag og Finnmark. I Trøndelag og Finnmark er økningen noen steder opp mot 75 og 100%, ellers er den stort sett mellom 0 og 25%. På østlandet, noen steder på vestlandet, Nordland og Troms modelleres det en reduksjon. Stort sett mellom 0 og 25%, men nord i Norland er reduksjon opp mot 75%.

For oktober er det store kontraster, med enten stor økningen eller kraftig reduksjon. På vestlandet og nordover langs mørekysten og mot Trøndelag modelleres det en reduksjon på opp mot 75%. I Troms og Finnmark er reduksjonen også kraftig med opp mot 100% flere steder. Øst i sørnorge, inn i landet, nord i Trøndelag og Nordland modelleres det kraftig økning, med flere områder opp mot 100%. Langs kysten av sørnorge er det riktignok små endringer, men verdier stort sett mellom 25% reduksjon og 25% økning.



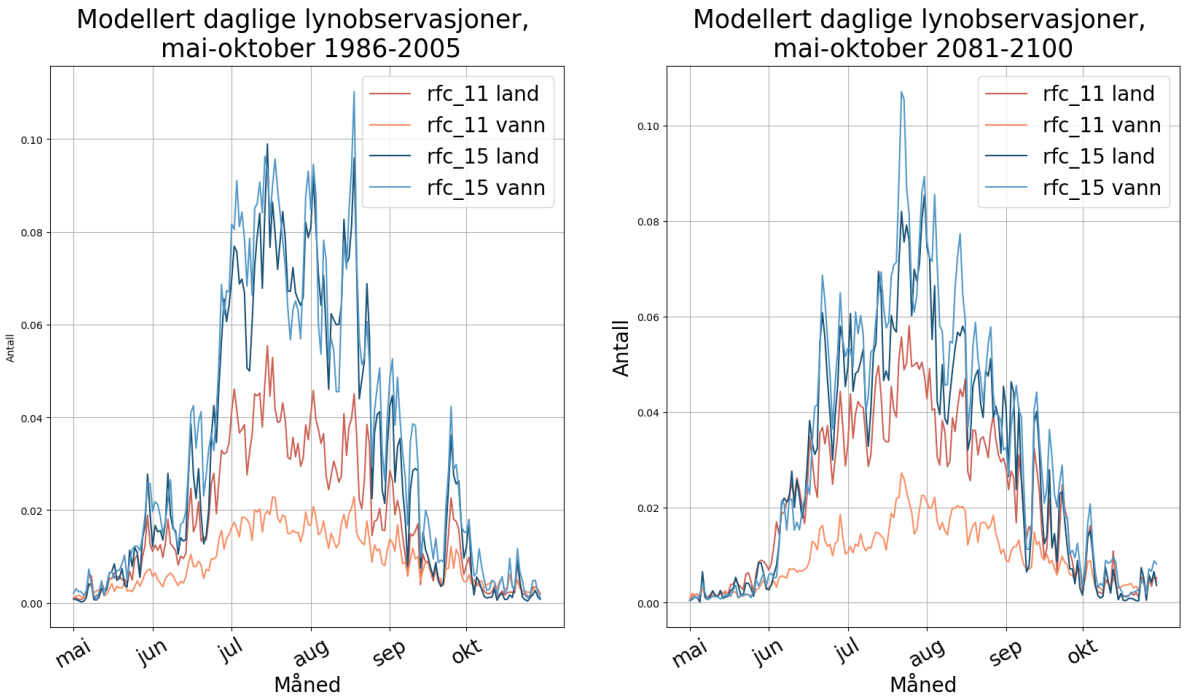
Figur 4.23: Resultater fra modell rfc_15. Figuren viser prosent endring i lynfrekvens per tidsenhet fra historisk til fremtidig periode. Endringene vises for månedene mai, juni, juli, august, september og oktober.

4.4.3 Sammenligning rfc_11 og rfc_15

I figur 4.24 vises den modellerte daglige gjennomsnittsfrekvensen for historisk periode til venstre, og fremtidig periode til høyre. Begge viser frekvensen over land og vann separat, og for både rfc_11 (rødtoner) og rfc_15 (blåtoner). For den historiske perioden modellerer rfc_15 en ganske lik frekvens for både land og vann, begge en god del høyere enn for rfc_11. Rfc_11 viser en god del lavere for vann enn for land. Begge viser høyest frekvens i juli og august, og lavest i oktober.

Fir den fremtidige perioden modellerer rfc_15 også ganske lik frekvens over land og vann, men ikke så mye høyere enn rfc_11. Rfc_11 modellerer adskillig lavere frekvens over vann enn over land. Her er det også høyest frekvens i juli og august, men juni er ikke så langt unna august. Nå er det minst aktivitet i mai.

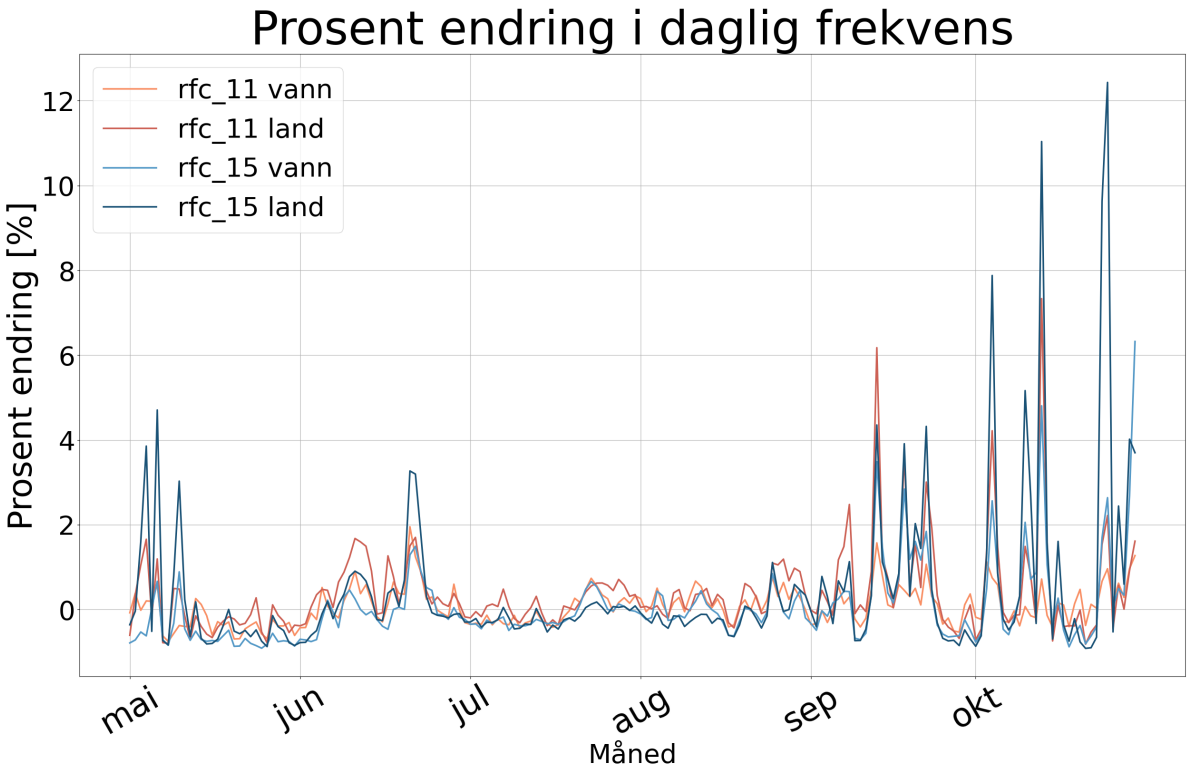
I figur 4.25 vises prosentvis endring fra historisk til fremtidig periode i den daglige gjennomsnittlige lynfrekvensen for hele området. rfc_11 vises i rødtoner, der mørkeste er og land og lyseste er over vann, og rfc_15 vises i blåtoner der mørkeste også er over land og lyseste er over vann.



Figur 4.24: Gjennomsnittlig modellert daglig lynfrekvens for månedene mai-oktober. Summen er over hele utstrekningen til kartutsnittet. Til venstre vises historisk periode, og til høyre vises fremtidig periode. Rfc_11 for vann vises i lys rød, og for land i mørk rød. Rfc_15 for vann vises i lys blå, og for land i mørk blå.

Modellene ligger ganske tett opp mot hverandre hele veien. Når modellene viser økning viser rfc_15 ofte mer dramatisk økning enn rfc_11, men ligger ellers ganske jevnt under. Begge ligger tidvis ganske tett rundt 0, men rfc_11 ligger ofte litt over, mens rfc_15 oftere ligger litt under.

I mai viser begge modellene er økning i starten av måneden, men mot slutten ligger begge litt under. I juni vise begge modellene litt nedgang i starten, men så en gradvis økning, før endringen går mot null igjen. I både juni og august viser begge modellene liten endring, men litt større svingninger mot slutten av august. I september blir det enda større svingninger, men noen store topper opp mot 4 og 6%, men en nedgang de siste dagene i måneden. I oktober modelleres de største svingningene. Spesielt rfc_15 viser flere store topper, den høyeste over 12%.



Figur 4.25: Prosent endring i daglig lynfrekvens fra den historiske til den fremtidige perioden for månedene mai-oktober. Rfc_11 for vann vises i lys rød, og for land i mørk rød. Rfc_15 for vann vises i lys blå, og for land i mørk blå.

5. Diskusjon

I dette kapittelet diskuteres resultatene. Kapitellet følger samme struktur som Resultater, og man kan gå til tilsvarende delkapittel for å se hvilke resultater som blir diskutert.

5.1 Utforskning av parametere

I dette delkapittelet diskuteres resultatene rundt utforskning av parametere.

5.1.1 Korrelasjons plot.

Korrelasjonsplottet for de kontinuerlige parameterne til treningsdataene er presentert i figur 4.1. De fleste parameterne er lite korrelert. Dermed er det mulig å bruke flest mulig i maskinlæring

At isotermene er mest korrelert gir veldig mening ettersom temperaturen i atmosfæren stort sett endrer seg kontinuerlig med trykknivå. At parameterne ikke er helt lineære tyder på at det ikke er en helt lineær sammenheng. Dette kommer av at lapseraten varierer, og det kan være interessant å ha dem med videre. Det kan for eksempel være interessant å se på om hvor stor avstand det er mellom dem har en betydning. Det kan også være interessant å sette isotermene i sammenheng med andre parametere, og se om funksjonsteknikk kan ha en innvirkning på resultatene.

At det er stor negativ korrelasjon mellom CAPE og CIN er heller ikke overraskende. I en stabil atmosfære er CIN høy og CAPE lav, og man forventer lite konveksjon. I en ustabil atmosfære er oftest CIN lav og CAPE høy, og man forventer mer konveksjon.

5.1.2 Sannsynlighetstetthetsplot

I figurene presentert i delkapittel 4.1.2 vises sannsynlighetstetthetsplottene for de kontinuerlige parameterne. Sannsynlighetstetthetplot egner seg godt til å se på dette datasettet siden det er såpass ubalansert, med veldig liten andel lyn. I disse plottene blir verdiene skalert etter hvor mye data det er totalt, og man kan sammenligne hele

datasettet og lynobservasjonene selv om det er veldig mange flere observasjoner totalt enn det er lyn.

Blå linje viser fordelingen av hele datasettet, mens den rød viser for de datapunktene som har lynobservasjon. Datasett uten lynobservasjon ble ikke tatt med siden lyn utgjør en såpass liten del av datasettet at det var tilsvarende som for hele datasettet. I plottetene til venstre er det normal skala, men til høyre er det semi-logaritmisk skala. Begge skalaene er tatt med fordi på den formale skalaen kan man lettere se hvordan verdiene fordeler seg, mens med den semi-logaritmiske skalaen får man tydeligere frem forskjeller mellom hele datasettet og lynobservasjoner.

For CAPE er det tydelig flest verdier nærmest 0, men den strekker seg opp mot 5000 Jkg^{-1}. På den semi-logaritmiske skalaen kommer det frem at lyn har relativt sett flere høye verdier for CAPE enn resten av datasettet. Dette gir mening ettersom lyn ofte blir satt i sammenheng med konvektiv aktivtet. Det betyr ikke nødvendigvis at høy CAPE gir lyn, men at når det er lyn, er det ofte høy CAPE.

For CIN ser man at det er flest verdier rundt 0 og 900 Jkg^{-1}. At det er mange verdier på 900 Jkg^{-1} kommer av at det er denne verdien som erstattet NaN-verdier i datasettet. Med den semi-logaritmiske skalaen kommer det frem et sprang i datasettet ved ca 350 Jkg^{-1}. Hvor dette kommer fra er usikkert, men dukket opp etter at dataene ble glattet ut med gridpp. Det som også kommer frem er at lavere verdier av CIN gir høyere sannsynlighetstetthet enn høye verdier med CIN. Dette gir mening ettersom lav CIN øker muligheten for konvektiv aktivitet, mens høy CIN forhindrer at luftpakker stiger i atmosfæren.

For RH ser man at verdiene fordeler seg mellom 0 og 1.2. Når RH er over en betyr det at lufta er overmetta. At høyere RH gir høyere sannsynlighetstetthet for lyn kommer frem i begge figurene for RH. Ettersom lyn ofte assosieres med hydrometeorer gir det mening at man må ha en viss fuktighet og mulighet for dannelse av hydrometeorer for å få lyn. Her er RH riktignok begrenset til isobaren ved 700hPa, og det kunne vært interessant å se på hvilken effekt det har hvis man ser på RH knyttet til en isotermer i stedet.

For w fordeler verdiene seg mellom 0 og 12 ms^{-1}, med flest verdier rundt 0. I figuren med semi-logaritmisk skala kommer det frem at lyn har høyere sannsynlighetstetthet når w er høy. Dette gir mening ettersom ei luftpakke med høyere vertikal hastighet når høyere i atmosfæren, og hvis det er hydrometeorer i luftpakka vil disse både bli avkjølt og anta gunstige faser, og bevege seg og ha mulighet for å utveksle ladninger. Stor vertikal hastighet sørger også for at ladede partikler blir separert, og det kan bygge seg opp en spenningsforskjell.

For 0-isotermen er det en tydelig normalfordeling, for både hele datasettet og for

lynobservasjoner, som kuttes av ved 1000 hPa. Det er også noen verdier ved 1500 hPa. Verdiene ved 1500 hPa kommer av at dette er verdien som erstattet NaN-verdier for 0-isotermen. Denne verdien ble valgt noe tilfeldig. Den er høyere enn hvor dataene kuttes av, men ikke veldig mye høyere. NaN-verdiene henger sammen med at datasettet kuttes ved 1000 hPa. Bakketrykket ligger normalt rundt dette, så hvis det er minusgrader ved bakken vil det ikke være mulig å måle trykket til den siden den ligger under bakken. I disse plottene gjelder det ikke mange tilfeller, noe som nok kommer av at det kun er mai-oktober og at det typsik er få tilfeller av minusgrader ved bakken da. Lynobservasjonene er normaltfordelt, men forskjøvet til venstre for resten av datasettet. Dette tyder på at for at lyn skal skje er det en fordel at isotermen har en viss høyde opp i atmosfæren. Spesielt ved 800 hPa er sannsynlighetstettheten høy. Hvorfor er ikke undersøkt, men det kan være interessant å se på 800 hPa isobaren i sammenheng med for eksempel RH for å se om det er noen sammenheng. Med den semi-logaritmiske skalaen kommer det også frem at det er noe tilfeller av lyn når trykket er satt til 1500 hPa, men det er relativt få.

Figurene for -15-isotermen ligner 0-isotermen, men fordeler seg mellom 500 og 950 hPa, og har ikke en avkutting av verdier ved 1000 hPa eller verdier ved 1500 hPa. Det vil si at det ikke var noen tilfeller i perioden som er undersøkt hvor -15-isotermen var ved bakkenivå eller lavere. Lynobservasjonene er også her normalfordelt og forskjøvet til venstre. Her er toppen rundt 620 hPa, noe høyere opp i atmosfæren enn for 0-isotermen. Dette er forventet ettersom temperaturen synker gradiv oppover, og -15-isotermen vil ha et lavere trykk enn 0-isotermen. Med den semi-logaritmiske skalaen kommer det frem noen tilfeller av lyn når trykket er på 890 hPa, og som er separert fra resten av lynobservasjonene. Det kunne vært interessant å se nermere på disse og om disse henger sammen med observasjonene som er gjort når 0-isotermen er under bakkenivå.

For clivi fordeler verdiene seg mellom 0 og 0.5. Det er flest verdier rundt 0, og man må se på den semi-logaritmiske skalaen for å hente mer info om hvordan lynobservasjonene er i forhold til resten av datasettet. Med semi-logaritmisk skala kan man se at høyere clivi gir høyere sannsynlighetstetthet for lyn enn lavere, noe som er som forventet: Høyere tetthet av hydrometeorer gir større mulighet for elektrifisering og lyn.

5.1.3 Multipliserte parametere

I delkapittel 4.1.3 vises 0-isotermen multiplisert med de andre parameterne, etter at de er normalisert til verdier mellom 0 og 1. Ikke alle figurene får frem ny info. For CAPE, w og clivi er det lite som egentlig endrer seg fra figurene der de presenteres alene.

I figur a) hvor CIN og 0-isotermen er multiplisert kommer det frem en ny fordeling, og datane blir tilsynelatende separert i to. En bolk mellom 0 og 0.5, og en på 1. Verdiene

ved 1 kommer nok av at både CIN og 0-isotermen har maksverdier som er substituert inn for NaN, og at det er såpass mange verdier på 1 kan tyde på at det ofte er en sammenheng mellom når de to parameterne er NaN. For verdiene mellom 0 og 0.5 er lynobservasjonene forskjøvet noe til venstre for det totale datasettet. At kombinasjonen lav verdi for CIN og lavt trykk for isotermen gir høyere sannsynlighetstetthet for lyn gir noe mening. Med lav CIN vil luftpakka ha mulighet til å stige, og toppen for lynaktivitet når man så på 0-isotermen var 800 hPa -et stykke opp i atmosfæren. Det kan være det som gir relativt høyere sannsynlighetstetthet for lyn rundt 0.1 i figuren for multipliserte parametre.

I figur d) hvor RH og 0-isotermen er multiplisert blir sannsynlighetstettheten for lyn separert i to. Den ene bolken er mellom 0 og 0.3, med en topp rundt 0.1, og den andre med en topp rundt 0.8. Sannsynlighetstettheten for hele datasettet er også litt merkelig, med en topp mellom 0 og 0.3, så ganske stabile verdier, frem til en liten topp ved 0.8. RH er ved 700hPa, mens isotermen endrer trykk, og dermed avstand til hvor RH er. De fleste av lynobservasjonene er når både RH og 0-isotermen er i det nedre sjiktet. Det er noe uventet at det er høyere sannsynlighetstetthet for lyn når RH er såpass lav, og det kunne vært interessant å se nærmere på.

I figur e), hvor -15-isotermen og 0-isotermen er multiplisert blir dataene ganske annerledes fordelt enn når isotermene plottes alene. Når produktet er lavt blir sannsynlighetstettheten til lyn relativt sett en del høyere enn for resten av dataene. Ved noen verdier er det også ikke noen sannsynlighet for lyn. Dette er rundt 0.4-0-5, og 0.9-1. At det er lite lyn når begge isotermer er lavt nede gir mening, men at det er lite når produktet er mellom 0.4 og 0.5 er litt merkeligere. Hva man kan hente fra dette plottet er litt usikkert, men noe som hadde vært mer interessant er å se på om avstanden mellom de to isotermene har noe å si for lynaktivteten.

Å multiplisere sammen parametere fikk i noen tilfeller frem nye fordelinger, og kanskje muligheten til å hente ut ny informasjon. Flere av kombinasjonene av parametere ble prøvd ut, og resultatene av dette diskuteres senere.

5.1.4 Lyndata

I delkapittel 4.1.4 vises resultatene for utforskning av lyndata. I figur 4.10 ser man tydelig at det er mest lyn i mai-oktober. Dette stemmer godt overens med hva tidligere studier har funnet (Køltzow mfl., 2018, Midtbø mfl., 2011), men lav registrert aktivitet resten av året kan også komme av manglende registrering av lynhendelser. Tidligere studier har funnet ut at det meste av vinterlynet skjer på vestlandet, og siden sensornettverket ikke har så god dekning er det ikke sikkert lynene blir registrert. I denne oppgaven ble det valgt å fokusere på mai-oktober blant annet på grunn av dette. Det er også denne

perioden det er flest lynobservasjoner og mest data å jobbe med.

I figur 4.11 vises gjennomsnittlig antall lyn, og gjennomsnittlig frekvens per tidsenhet for hver gridcelle i et tidsintervall på 3 timer. I figuren til venstre er det enkelte steder med ganske høyt gjennomsnitt som ikke vises i plottet til venstre. Dette gjelder spesielt Nordmøre/Trøndelag og Finnmarksvidda. Siden plottet til venstre tar antall lyn i betraktning kan enkelthendelser med mange lyn på kort tid har stor innvirkning. I figuren til høyre er det kun om det er en lynhendelse eller ikke som blir tatt i betraktning. Figuren til venstre kan derfor si mer om intensiteten i lynhendelsene, mens figuren til høyre sier mer om hvor det oftest skjer lyn. For å forenkle oppgaven og dataene er det frekvensen i lynhendelser som blir brukt. Som kartene i figur 4.11 viser får man dermed ikke frem informasjon om intensiteten i lyn, men man får likevel en viss aning på hvor det er mest.

I figure 4.12 ser man når på døgnet det er mest lynhendelser. Tidspunkt er et interessant parameter å ta med, siden det er en tydelig sammenheng med når på dagen det lynes mest. Hvilken dag i året ble en stund vurdert å ha med som parameter, men ble droppet siden noe av det som kan være interessant er hvorvidt lynaktiviteten endrer seg i løpet av året, og med dag i året som et parameter kan dette overtilpasse modellen.

5.2 Preprosesseringa

I dette delkapittelet diskuteres hvordan resultatene fra og underveis i preprosesseringa ble, hvorfor enkelte avgjørelser ble tatt og hvilken betydning de hadde.

5.2.1 CDO og gridpp

Del 1 av preprosesseringa ble gjort med CDO og gridpp, og er ganske effektivt når man lærer å bruke det. For å sette sammen et datasett, utføre enkle utregninger og maskere ut områder i kartet var spesielt CDO et kraftig verktøy. Å kjøre hele del 1 av preprosesseringa, fra P1 til D2 (se figur 3.2, tok i underkant av et døgn per år, men prosessene kunne linkes etter hverandre og stå og gå av seg selv. Mesteparten av tida gikk til gridpp, som glattet ut enkelte av parameterne. Hvor stor innvirkning det hadde på parameterne ble dessverre ikke undersøkt i dybden, men visualisert som i figur 4.13. Det jevnet i alle fall ut verdiene, og hensikten er at det skal fjerne store lokale variasjoner.

Videre ble reduksjon av datasettet gjort med python. Det å redusere til kun mai-oktober halverte størrelsen. I tabell 4.1 ble dette gjort for kolonna A til B, men det ble også lagt til to nye parametere som gjorde at filstørrelsen ble litt mer enn halvparten. Fra B til C ble det gjort store reduksjoner. NaN verdier ble fjernet, indeksen forenklet, og flere av datapunktene hvor CIN = 900 Jkg^{-1} fjernet. NaN-verdiene kunne fjernes i C siden

dette datasettet kun skulle brukes til å trene og teste modeller. Derfor ble også indeksen forenklet fra å inneholde informasjon om posisjon og tid, til å kunn være et nummer. Siden CIN-verdier på 900Jkg^{-1} var en stor overvekt i datasettet, ble det omtrent halvvert ved å fjerne 75% av disse verdiene. Det utgjorde fortsatt en stor del av datasettet, for noen år opp mot halvparten, men det trengs ikke en så stor overvekt av samme verdi for et parameter. Resultatet ble et datasett som var av håndterbar størrelse og klar for bruk i maskinlæring.

5.3 Maskinlæringsmodeller

Byggingen av maskinlæringsmodellene besto i stor grad å prøve forskjellige kombinasjoner av parametere, for så å videreutvikle de som virket mest lovende.

5.3.1 Tidlig utprøving av modeller

Det var mange forskjellige kombinasjoner av parametere som ble prøvd ut, med varierende suksess. Det ble tydelig veldig fort at rfc-modeller gjorde det adskillig bedre enn logreg-modeller. Logreg_0 ble derfor kun tatt med videre som en basemodell.

Det ble tydelig at flere parametere ga bedre resultater, noe som er som forventet. Flere parametere gir mer data å hente. De multipliserte parameterne ga et interessant resultat. Det gjorde at sannsynligheten for lyn generelt ble redusert, og påvirket spesielt f1-scoren. Grenseverdien her var standardverdien på 0.5, noe som betydde at med lavere sannsynlighet ble det mange flere FP. ROC-AUC ble derfor vektlagt mer. Modellene rfc_3, rfc_4, rfc_10 og rfc_15 ble tatt med videre fordi de hadde høye verifikasjonsverdier. Rfc_11 ble tatt med videre av nysgjerrighet, og viste seg å prestere veldig bra på blant annet reliabilitetsdiagram. Selv om rfc_2, rfc_12 og rfc_14 også hadde ganske bra scores ble de forkastet siden de ikke inneholdt land/vann-maske som et paramter, og det var ønskelig å kunne skille på dette. Det kunne vært interessant å jobbe videre med disse også, men ikke alle kombinasjoner av alle parametere kunne taes med.

5.3.2 Videre utvikling av enkelte modeller

Under videre utvikling av modellene ble de først trent og testet på et større datasett. Dette var for å se hvordan dette påvirket ROC-AUC-scoren, og om de overtilpasset seg til dataene. Det var liten antydning til overtilpassing, men ROC-AUC-scorene til rfc-modellene ble mye likere enn når de kun ble trent på ett år. De ble også stort sett lavere. Dette kan tyde på at den store variasjonen mellom få år gjør det vanskeligere å lage treffsikre modeller, og at hvilke parametere man har med og ikke har mindre å si

for tilpasning av modellen.

Grenseverdiene til modellene uten multipliserte parametere var adskillig høyere enn for de med multipliserte parametere. Dette tyder på at modellene uten multipliserte parametere predikerer en høyere sannsynlighet for lyn enn de som har multipliserte parametere. Bortsett fra logreg_0 har samtlige modeller lavere grenseverdi for vann enn for land. Dette tyder på at det er forskjeller på frekvensen i lyn over land og vann.

Ingen av modellene utpekte seg som spesielt gode til å rangere årene. Alle traff riktig på tre av årene for lynfrekvens over land, men over vann var det flere som gjorde det dårligere, samtlige modeller med multipliserte parametere. At det er forskjell over vann og land tyder igjen på at det er forskjeller her, og to forskjellige modeller som ser på land og vann hver for seg kunne vært gunstig.

F1-scoren til alle modellene ble ganske lik. For de modellene med multipliserte parametere gikk den ned i forhold til når de ble trent og testet på kun 2015, men for de med multipliserte parametere gikk den opp. For f1-scoren i tabell 4.3 var grenseverdien standard 0.5, i tillegg til at det var et mindre datasett, så begge disse faktorene kan være med på å påvirke verdien. Spesielt grenseverdiene. Det kan tyde på at de nye grenseverdiene blir litt høye for de uten multipliserte parametere, og at de presterer bedre under mer standard forhold. Men, for de med multipliserte parametere ser det ut til å være mer gunstig med de nye grenseverdi.

5.3.3 Reliabilitetsdiagram

I reliabilitetsdiagrammene ser man enda tydeligere at rfc_10, som ikke har multipliserte parametere, overestimerer kraftig sannsynligheten for lyn, sammenlignet med den observerte frekvensen. Rfc_11 og rfc_15 overestimerer også, men ligger litt tettere opp mot diagonalen. Det tyder på at de ikke overestimerer fullt så mye, og tyder igjen på at multipliserte parametere generelt senker den generelle modellerte sannsynligheten for lyn.

I valg av utvelgelse av modeller videre ble reliabilitetsdiagrammene vektlagt en del, samt at modeller med multipliserte parametere hadde en lavere modellert sannsynlighet for lyn. Derfor ble rfc_11 og rfc_15 brukt til å modellere lynaktivitet videre. Rfc_3 ble ikke brukt videre ettersom det hadde litt lavere ROC enn rfc_11 og rfc_15, og eneste forskjell var at clivi var med som et parameter. Clivi ble også rangert som minst betydningsfullt, og reduserte datamengden litt. Rfc_11 ble også tatt med videre for å se hvordan redusering av antall parametere påvirket resultatene.

5.3.4 Visualisering av prediksjoner til modeller

I delkapittelet 4.3.4 ble resultatene av predikasjonene gjort av rfc_11 og rfc_15 for årene 2014-2018 visualisert i kart. I b) plottet til både figur 4.18 og figur 4.19 er rfc_11 og rfc_15 ganske samstemte om hvor det er lynaktivitet. I c) plottene kan man også se at de samsvarer ganske bra med tanke på hvor de over- og underestimerer. Ut i fra disse figurene ser det ikke ut til at det har en stor betydning om CIN og -15-isotermene, som brukes i rfc_15, men ikke rfc_11, er med eller ikke.

Begge modellene treffer ganske bra med hvor det er observert høy og lav lynfrekvens. I de observerta dataene ser det ut til at det går et slags skille ved fjellene i sørnorge der det er mer lyn østover. Dette vises også i de modellerte dataene, selv om det er noe forskjøvet vestover. For sørøst-Sverige overestimeres det kraftig, og det treffer ikke helt i Finnland heller. Av landene som er med i figuren er det Norge som blir best modellert. Det er kraftig overestimering i Sørnorge, og en del underestimering på Finnmarksvidda, men ellers er det relativt lite avvik for resten av fastlandet.

En stor svakhet i disse predikasjonene er at dataene som er brukt til å trene modellen, er et utdrag fra samme datasett som blir brukt til å predikere og visualisere modellert lynaktivteten. Dette ikke ideelt, men gir likevel en mulighet til å se hvordan modellene som brukes videre presterer på dataene som brukes i treningen.

5.4 Modellerte endringer i lynaktivitet

Både rfc_11 og rfc_15 ble brukt til å predikere endringer i lynaktivitet. Ingen av dem hadde perfekte ytelsesmål, men de hadde bedre enn mange andre modeller, og er derfor hva som brukes for å se på endringer i lynaktivitet. Her diskuteres både modelleringen av histrisk aktivitet og fremtidig aktivitet gjort av begge modeller, prosent endring i aktivitet, og også forskjeller mellom de to modellene.

5.4.1 rf_11

Lynaktivitet modellert av rfc_11 for den historiske perioden har likheter med perioden man har observerte lyn, men også forskjeller. En forskjell er at den modellerte frekvensen er noe høyere enn den observerte, og at det er litt forskjell i hvor det modelleres og observeres høyest aktivtet. Observasjonsdataene viser størst aktivitet på østlandet og langs vestkysten av Sverige, mens den modellerte frekvens per tidsenhet er høyest på sørvest-landet og helt sør i Sverige. Den historiske perioden går fra 1986-2005, og det kan være at lynaktiviteten har endret seg siden da, men at det er så stor endring er lite sannsynlig. Det modelleres også en del høyere frekvens per tidsenhet på Vestlandet og Nordland, men tendensene om en lavere frekvens per tidsenhet jo lengre nord man

kommer er der. Det ser ikke ut til at modelleringa er helt feil, men at det er noen større elementer som tyder på at den ikke er særlig presis. Spesielt at at det ikke er et tydelig mellom øst og vest i Sørnorge, slik som det er for de observerte dataene, er noe som taler for at den ikke treffer helt.

For den fremtidige perioden er det også mest aktivitet på sør og sørvest-landet, og gradvis mindre nordover. Frekvensen ser ut til å være noe økende i forhold til den historiske perioden. Om den modellerte fremtidige frekvensen stemmer er det vanskeligere å si noe om, men flere studier har vist at lynaktiviteten kan øke (Rädler mfl., 2019, Finney mfl., 2018, Kahraman mfl., 2022).

I figuren hvor prosent endring vises, ser man at det jevnt over er en økning. Det er ikke mye, for det meste mellom 0 og 25%, men noen steder er det opptil 50% økning. For å se på hvorfor det er en økning må man se nærmere på de forskjellige parameterne og hvordan de endrer seg.

I figur 4.21 hvor man kan se modellert prosent endring for de forskjellige månedene kommer det frem at det er store variasjoner fra måned til måned. Midt på sommeren er det relativt lite endringer, men på starten og slutten ser det ut til at det kan bli store endringer. For å se på hvorfor det er såpass store forskjeller fra måned til måned bør man se på parameterne som brukes i modelleringa. Ut i fra denne figuren kan man bare si at det ser ut til at det endringer. En måned som er spesielt verdt å merke seg er oktober. Her er det et tydelig skille mellom øst og vest i Sørnorge, med skille omtrent hvor man ser det på de observerte dataene. At det modelleres såpass stor økning øst, mens det modelleres kraftig reduksjon i vest burde sees nærmere på.

5.4.2 rfc_15

Den modellerte lynfrekvensen for rfc_15 er preget av at frekvensen over vann er en god del høyere enn over land, og at over land er den høyest i sør/sørøst-Norge, og sør i Sverige og Finnland. Den går også her gradvis nedover jo lengre nord man kommer. Den er generelt en del høyere enn den observerte frekvensen, og har noen av de samme tendensene, men her mangler det også et skille mellom øst og vest i Sørnorge. Mye av det samme kan sies som for rfc_11; lynfrekvensen kan ha endret seg, men det er lite sannsynlig at den har endret seg så mye.

For den fremtidige perioden er det samme gradvise nedgang i frekvens per tidsenhet nordover, men også en nedgang i frekvens generelt.

Dette kommer tydelig frem i plottet hvor prosent endring vises. Her ser man at det over størstedelen av landet modelleres en reduksjon på mellom 0 og 25%, og noen steder i Nord-Norge er reduskjon på opptil 50%. Hvorfor det her vises en reduksjon er usikkert,

men rfc_15 inkluderer blant annet -15-isotermen, som ofte assosieres med elektrifisering av partikler (Kahraman mfl., 2022), og med en varmere atmosfære kan det være den blir såpass høyt at hydrometeorene ikke vil nå opp, selv om det blir mer konvektivitet.

I figur 4.23 vises modellert prosent endring for mai-oktober. Det er stor variasjon mellom månedene. Starten har både kraftig reduksjon og noe økning, mens midt på sommeren er det en litt reduksjon over hele kartet. For september og oktober er det store forskjeller, med noen områder med kraftig økning og andre med mye reduksjon. Spesielt oktober har store forskjeller, her også med et skille mellom øst og vest i Sørnorge på samme måte som oktober fra rfc_11.

5.4.3 Sammenligning rfc_11 og rfc_15

Hvilke av rfc_11 og rfc_15 som gir den beste predikasjonen på endring av lynaktivitet er vanskelig å si. De har omtrent like bra verifikasjonsverdier, men de gir motsatt resultat på om hvordan den totale endringen i lynaktivitet blir. Siden forskjellen på oppbyggingen av modellene er hvilke parametere som er med og ikke, viser dette at hvilke parametere man ser på har mye å si for hvordan endringen blir. Det er også flere parametere som ikke har vært med i denne oppgaven som er aktuelle å se på, spesielt parametere som kan si mer om hvilke hydrometeorer som er i regionene av atmosfæren hvor elektrifisering normalt finner sted.

Figurene presentert i delkapittel 4.4.3 sammenligner de gjennomsnittlige daglige frekvensene for rfc_11 og rfc_15. I figur 4.24 ser man at rfc_15 estimerer en god del høyere frekvens per tidsenhet enn rfc_11 for den historiske perioden, men at de er nermere hverandre for den fremtidige perioden. Dette var også synlig på kartene i delkapittel 4.4.1 og 4.4.2, men i figur 4.24 er frekvensen for gjennomsnittlig lynaktivitet i både rom og tid. En forskjell på modellene er at rfc_15 modellerer ganske lik frekvens per tidsenhet over land og vann, både for historisk og fremtidig periode. Mens rfc_11 modellerer adskillig lavere frekvens per tidsenhet over vann, omtrent halvparten av frekvensen over land.

Hvilke måneder det er mest og minst aktivitet stemmer godt overens med observerte data, men at det er såpass store svingninger er noe uventet. Siden verdiene er for et gjennomsnitt over 20 år kunne man forventet mindre dramatiske topper, og en glattere kurve. At det på samme dag i 20 år er en markant høyere frekvens per tidsenhet enn dager før eller etter er lite sannsynlig, men det kan være enkeltdager med veldig høy eller lav frekvens som påvirker gjennomsnittet over hele perioden såpass mye at det synes. Om dette er tilfellet må resultatene undersøkes nærmere.

I figur 4.25 ser man prosent endring i den daglige, gjennomsnittlige frekvensen for de to

modellene. Det er som er mest interessant med denne figuren er at begge modellene har ganske like mønstre. De samsvarer ganske bra med når endringen går opp eller ned, og når det er store svingninger. De er riktignok litt forskjøvet i forhold til hverandre, med at resultatene fra rfc_15 har større svingninger og gir en høyere frekvens per tidsenhet når frekvensen øker, og en lavere frekvens per tidsenhet når den minker eller er ganske stabil. Siden modellene er litt uenige er det vanskelig å si hvor mye man kan forvente at frekvensen endrer seg, men at de samsvarer ganske bra på noen områder styrker disse betraktningene. At begge modellerer noe større aktivitet i mai og juni, tyder på at lynaktiviteten vil øke på slutten av våren/starten av sommeren. Ikke drastisk, men noe. At begge modellene også viser lite endringer i juli og august tyder på at man kan forvente at lynaktiviteten blir ganske stabil. Kanskje går den litt ned, kanskje går den litt opp, men det er ikke drastiske endringer. De drastiske endringene dukker mer sannsynlig opp i september og oktober. Her gir begge modellene tall som tyder på at man kan forvente økt aktivitet. Det er ikke jevnt fordelt over hele perioden, noen dager viser også en nedgang, men det er flere dominerende topper. Dette kan tyde på at det ikke blir mer aktivitet jevnt over, men at man kanskje kan få enkelte dager med stor sannsynlighet for lyn, og at disse dagene har en stor innvirkning på den gjennomsnittlige daglige frekvensen. Dataene må riktignok undersøkes nermere for å si noe mer om dette.

6. Konklusjon

Motivasjonen til denne oppgaven var å bygge en modell ved hjelp av maskinlæring, lynobservasjoner og klimamodelldata, og å bruke denne modellen til å modellere endringer i lynaktiviteten i Norge for perioden 2081-2100. I denne seksjonen presenteres hovedfunnene gjort i oppgaven, i tillegg til potensielt videre arbeid.

6.0.1 Hovedfunn

Hovedfunnene i denne oppgaven kan summeres opp med:

- Å bruke klimamodelldata sammen med lynobservasjoner i maskinlæring viser potensiale, men modellene utviklet i denne oppgaven har mye rom for forbedring. Samtlige overestimerer lynfrekvensen per tidsenhet, og trenger kalibrering.
- RFC presterer bedre enn logistisk regresjon på dette datasettet.
- Å inkludere isotermer som parametre forbedrer modellene.
- Modellene presterer ulikt for områder som er over land eller vann. De presterer best over land.
- Hvilke parametere som inkluderes endrer hvorvidt det modelleres en økning eller reduksjon i lynaktivitet.
- Det er usikkert hvorvidt lynaktiviteten i Norge vil øke eller minke hvis man ser på hele perioden mai-oktober.
- Det er en fordel å se på kortere perioder, gjerne måned for måned, når man ser på endringer i lynaktivitet. Det er store variasjoner gjennom året, men modellene brukt til å modellere endring i lynaktiviteten viste størst endringer i mai, juni, september og oktober. Juli og august viser lite endringer, og modellene viser ulikt resultat på om det øker eller reduseres.
- Begge modellene modellerte en reduksjon av lynaktivitet i mai, bortsett fra på Nordmøre hvor det vises en kraftig økning.
- Begge modellene viser en økning i aktivitet i juni og september for størstedelen av landet, men i varierende grad.
- Begge modellene viser kraftig reduksjon på Vestlandet, men en kraftig økning på Østlandet i oktober.

6.1 Videre arbeid

Det er enda mye å oppdage innenfor dette temaet. Her er noen punkter med videre arbeid som kan være interessant å se nermere på:

- Datasettet med lynobservasjoner burde utvides til en lengre homogen periode.
- Endring i lynaktivitet for november-april bør undersøkes.
- Flere maskinlæringsmodeller kan undersøkes. Nevrale nettverk kan være spesielt interessant å undersøke, siden de har mulighet for å se på romlig utstrekning.
- Flere parametere burde undersøkes. Det å se på avstand mellom isotermer, og å inkludere flere parametere som beskriver patikler i atmosfæren er spesielt interessant.
- Lokale og regionale endringer burde undersøkes næremere.
- Endringer i inputparameterne brukt i modellene burde undersøkes nærmere for å undersøke hva som ligger bak endringen i lynaktivitet.

Referanser

Ávila, E. E., Bürgesser, R. E., Castellano, N. E., Collier, A. B., Compagnucci, R. H. og Hughes, A. R. W. (sep. 2010). Correlations between deep convection and lightning activity on a global scale. *Journal of Atmospheric and Solar-Terrestrial Physics* 72 (14): 1114–1121. DOI: 10.1016/j.jastp.2010.07.019.

Battaglioli, F., Groenemeijer, P., Púčik, T., Taszarek, M., Ulbrich, U. og Rust, H. (sep. 2023). Modelled multidecadal trends of lightning and (very) large hail in Europe and North America (1950–2021). EN. *Journal of Applied Meteorology and Climatology* -1 (aop). Publisher: American Meteorological Society Section: Journal of Applied Meteorology and Climatology. DOI: 10.1175/JAMC-D-22-0195.1.

Belušić, D., de Vries, H., Dobler, A., Landgren, O., Lind, P., Lindstedt, D., Pedersen, R. A., Sánchez-Perrino, J. C., Toivonen, E., van Ulft, B., Wang, F., Andrae, U., Batrak, Y., Kjellström, E., Lenderink, G., Nikulin, G., Pietikäinen, J.-P., Rodríguez-Camino, E., Samuelsson, P., van Meijgaard, E. og Wu, M. (mar. 2020). HCLIM38: a flexible regional climate model applicable for different climate zones from coarse to convection-permitting scales. English. *Geoscientific Model Development* 13 (3). Publisher: Copernicus GmbH: 1311–1333. DOI: 10.5194/gmd-13-1311-2020.

Bright, D., Wandishin, M. S., Jewell, R. E. og Weiss, S. (2004). A Physically Based Parameter for Lightning Prediction and its Calibration in Ensemble Forecasts. I: URL: https://www.semanticscholar.org/paper/A-Physically-Based-Parameter-for-Lightning-and-its-Bright-Wandishin/c21c3b44544f3513118ad2885e9c502b8e43905e (sjekket 17.01.2024).

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., VanderPlas, J., Joly, A., Holt, B. og Varoquaux, G. (2013). API design for machine learning software: experiences from the scikit-learn project. I: *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*: 108–122.

Calvin, K. mfl. (jul. 2023). *IPCC, 2023: Climate Change 2023: Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, H. Lee and J. Romero (eds.)]. IPCC, Geneva, Switzerland.* en. Tekn. rapp. Edition: First. Intergovernmental Panel on Climate Change (IPCC). DOI: 10.59327/IPCC/AR6-9789291691647.

Dong, G. og Liu, H. (mar. 2018). *Feature Engineering for Machine Learning and Data Analytics*. en. Google-Books-ID: 661SDwAAQBAJ. CRC Press.

Dwyer, J. R. og Uman, M. A. (jan. 2014). The physics of lightning. en. *Physics Reports* 534 (4): 147–241. DOI: 10.1016/j.physrep.2013.09.004.

Finney, D. L., Doherty, R. M., Wild, O., Stevenson, D. S., MacKenzie, I. A. og Blyth, A. M. (mar. 2018). A projected decrease in lightning under climate change. en. *Nature Climate*

Change 8 (3). Number: 3 Publisher: Nature Publishing Group: 210–213. DOI: 10.1038/s41558-018-0072-6.

Geng, Y.-a., Li, Q., Lin, T., Yao, W., Xu, L., Zheng, D., Zhou, X., Zheng, L., Lyu, W. og Zhang, Y. (2021). A deep learning framework for lightning forecasting with multi-source spatiotemporal data. en. *Quarterly Journal of the Royal Meteorological Society* 147 (741). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/qj.4167: 4048–4062. DOI: 10.1002/qj.4167.

Groenemeijer, Púcik, Tsonevsky, I. og Bechtold, P. (nov. 2019). *An Overview of Convective Available Potential Energy and Convective Inhibition provided by NWP models for operational forecasting*. Technical memorandum. DOI: 10.21957/q392hofrl.

Hanssen-Bauer, I., Førland, E., Haddeland, I., Hisdal, H., Mayer, S., Nesje, A., Nilsen, J., Sandven, S., Sandø, A., og B. Ådlandsvik, A. S., Andreassen, L., Beldring, S., Bjune, A., Breili, K., Dahl, C. A., Dyrrdal, A., Isaksen, K., Haakenstad, H., Haugen, J., Hygen, H., Langehaug, H., Lauritzen, S.-E., Lawrence, D., Melvold, K., Mezghani, A., Ravndal, O., Risebrobakken, B., Roald, L., Sande, H., Simpson, M., Skagseth, Ø., Skaugen, T., Skogen, M., Støren, E., Tveito, O. og Wong, W. (2015). *Klima i Norge 2100*. Meteorologisk institutt.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C. og Oliphant, T. E. (sep. 2020). Array programming with NumPy. *Nature* 585 (7825): 357–362. DOI: 10.1038/s41586-020-2649-2.

Holle, R. L. (okt. 2014). Some aspects of global lightning impacts. I: *2014 International Conference on Lightning Protection (ICLP)*: 1390–1395. DOI: 10.1109/ICLP.2014.6973348.

Holzworth, R. H., Brundell, J. B., McCarthy, M. P., Jacobson, A. R., Rodger, C. J. og Anderson, T. S. (2021). Lightning in the Arctic. en. *Geophysical Research Letters* 48 (7). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2020GL091366: e2020GL091366. DOI: 10.1029/2020GL091366.

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering* 9 (3): 90–95. DOI: 10.1109/MCSE.2007.55.

J, X. (sep. 2020). *A brief introduction to uncertainty calibration and reliability diagrams*. en. URL: https://towardsdatascience.com/introduction-to-reliability-diagrams-for-probability-calibration-ed785b3f5d44 (sjekket 08.03.2024).

Johns, R. H. (1992). Severe Local Storms. en. *W E AT H E R A N D F O R E C A S T I N G* 7.

Kahraman, A., Kendon, E. J., Fowler, H. J. og Wilkinson, J. M. (okt. 2022). Contrasting future lightning stories across Europe. en. *Environmental Research Letters* 17 (11). Publisher: IOP Publishing: 114023. DOI: 10.1088/1748-9326/ac9b78.

Kashinath, K., Mustafa, M., Albert, A., Wu, J.-L., Jiang, C., Esmaeilzadeh, S., Azizzadenesheli, K., Wang, R., Chattopadhyay, A., Singh, A., Manepalli, A., Chirila, D., Yu, R., Walters, R., White, B., Xiao, H., Tchelepi, H. A., Marcus, P., Anandkumar, A., Hassanzadeh, P. og Prabhat, n. (feb. 2021). Physics-informed machine learning: case studies for weather and climate modelling. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 379 (2194). Publisher: Royal Society: 20200093. DOI: 10.1098/rsta.2020.0093.

Køltzow, M., Dobler, A. og Eide, S. S. (2018). *Lightning in Norway under a future climate*. English. Tekn. rapp. 9.

Køltzow, M. A. Ø. (2023). Meteorologisk institutt. Privat kommunikasjon.

Lamb, D. og Verlinde, J. (2011). *Physics and Chemistry of Clouds*. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9780511976377.

Lind, P., Belušić, D., Christensen, O. B., Dobler, A., Kjellström, E., Landgren, O., Lindstedt, D., Matte, D., Pedersen, R. A., Toivonen, E. og Wang, F. (okt. 2020). Benefits and added value of convection-permitting climate modeling over Fenno-Scandinavia. en. *Climate Dynamics* 55 (7): 1893–1912. DOI: 10.1007/s00382-020-05359-3.

Lind, P., Belušić, D., Médus, E., Dobler, A., Pedersen, R. A., Wang, F., Matte, D., Kjellström, E., Landgren, O., Lindstedt, D., Christensen, O. B. og Christensen, J. H. (jul. 2023). Climate change information over Fenno-Scandinavia produced with a convection-permitting climate model. en. *Climate Dynamics* 61 (1): 519–541. DOI: 10.1007/s00382-022-06589-3.

McKinney, W. (2010). Data Structures for Statistical Computing in Python. I: *Proceedings of the 9th Python in Science Conference*. Red. av S. van der Walt og J. Millman: 56–61. DOI: 10.25080/Majora-92bf1922-00a.

MET (2024). *metno/gridpp: Software to post-process gridded weather forecasts*. URL: https://github.com/metno/gridpp/tree/master (sjekket 14.05.2024).

Midtbø, K. H., Haugen, J. E. og Køltzow, M. (2011). *Lynstudien Klimaendringenes betydning for forekomsten av lyn og tilpasningsbehov i kraftforsyningen*. Norwegian. Tekn. rapp.

NRK (jul. 2021). *Søstre døde av lynnedslag*. nb-NO. Section: nyheter. URL: https://www.nrk.no/nyheter/sostre-dode-av-lynnedslag-1.15564231 (sjekket 13.05.2024).

Oyj, V. (2015). *TLPTM SERIES USER'S GUIDE*. English. Vaisala Oyj.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. og Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12: 2825–2830.

Raschka, S. og Mirjalili, V. (2019). *Python Machine Learning: Machine Learning and Deep Learning with Python, Scikit-Learn, and TensorFlow 2, 3rd Edition*. en. Google-Books-ID: n1cjyAEACAAJ. Packt Publishing.

RealTek, N. (2024). *Kunstig intelligens ved REALTEK | NMBU*. nb. URL: https://www.nmbu.no/fakulteter/fakultet-realfag-og-teknologi/kunstig-intelligens-ved-realtek (sjekket 15.05.2024).

Rivrud, K. (aug. 2016). *Slik kunne lynet drepe 323 villrein*. nb-NO. Section: dk. URL: https://www.nrk.no/vestfoldogtelemark/slik-kunne-lynet-drepe-323-villrein-1.13110759 (sjekket 13.05.2024).

Romps, D. M., Charn, A. B., Holzworth, R. H., Lawrence, W. E., Molinari, J. og Vollaro, D. (2018). CAPE Times P Explains Lightning Over Land But Not the Land-Ocean Contrast. en. *Geophysical Research Letters* 45 (22). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2018GL080267: 12, 623–12, 630. DOI: 10.1029/2018GL080267.

Rädler, A. T., Groenemeijer, P. H., Faust, E., Sausen, R. og Púčik, T. (aug. 2019). Frequency of severe thunderstorms across Europe expected to increase in the 21st century due to rising instability. en. *npj Climate and Atmospheric Science* 2 (1). Number: 1 Publisher: Nature Publishing Group: 1–5. DOI: 10.1038/s41612-019-0083-7.

Salomonsen, M. (2024). Meteorologisk institutt. Privat kommunikasjon.

Sidselrud, L. F. (2024). Meteorologisk institutt. Privat kommunikasjon.

Soula, S. (des. 2012). Electrical Environment in a Storm Cloud. *Aerospace Lab* (5). Publisher: Alain Appriou: p. 1–10. URL: https://hal.science/hal-01184322 (sjekket 12.05.2024).

Stull, R. B. (2017). *Practical Meteorology: An Algebra-based Survey of Atmospheric Science*. en. Google-Books-ID: sjVfAAAACAAJ. Department of Earth, Ocean & Atmospheric Sciences, University of British Columbia.

Ushio, T., Heckman, S., Boccippio, D., Christian, H. og Kawasaki, Z.-I. (okt. 2001). A survey of thunderstorm flash rates compared to cloud top height using TRMM satellite data. *Journal of Geophysical Research* 106: 24089–24095. DOI: 10.1029/2001JD900233.

Yang, J., Wang, Z., Heymsfield, A. J. og French, J. R. (aug. 2016). Characteristics of vertical air motion in isolated convective clouds. English. *Atmospheric Chemistry and Physics* 16 (15). Publisher: Copernicus GmbH: 10159–10173. DOI: 10.5194/acp-16-10159-2016.

Yoshida, S., Morimoto, T., Ushio, T. og Kawasaki, Z. (2009). A fifth-power relationship for lightning activity from Tropical Rainfall Measuring Mission satellite observations. en. *Journal of Geophysical Research: Atmospheres* 114 (D9). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2008JD010370. DOI: 10.1029/2008JD010370.

Vedlegg A. Operasjonell lynalgoritme ved Meteorologisk insitutt

Den operasjonelle lynalgoritmen ble delt i forbindelse med denne oppgaven oktober 2023 av Morten Andreas Ødegaard Køltzow.

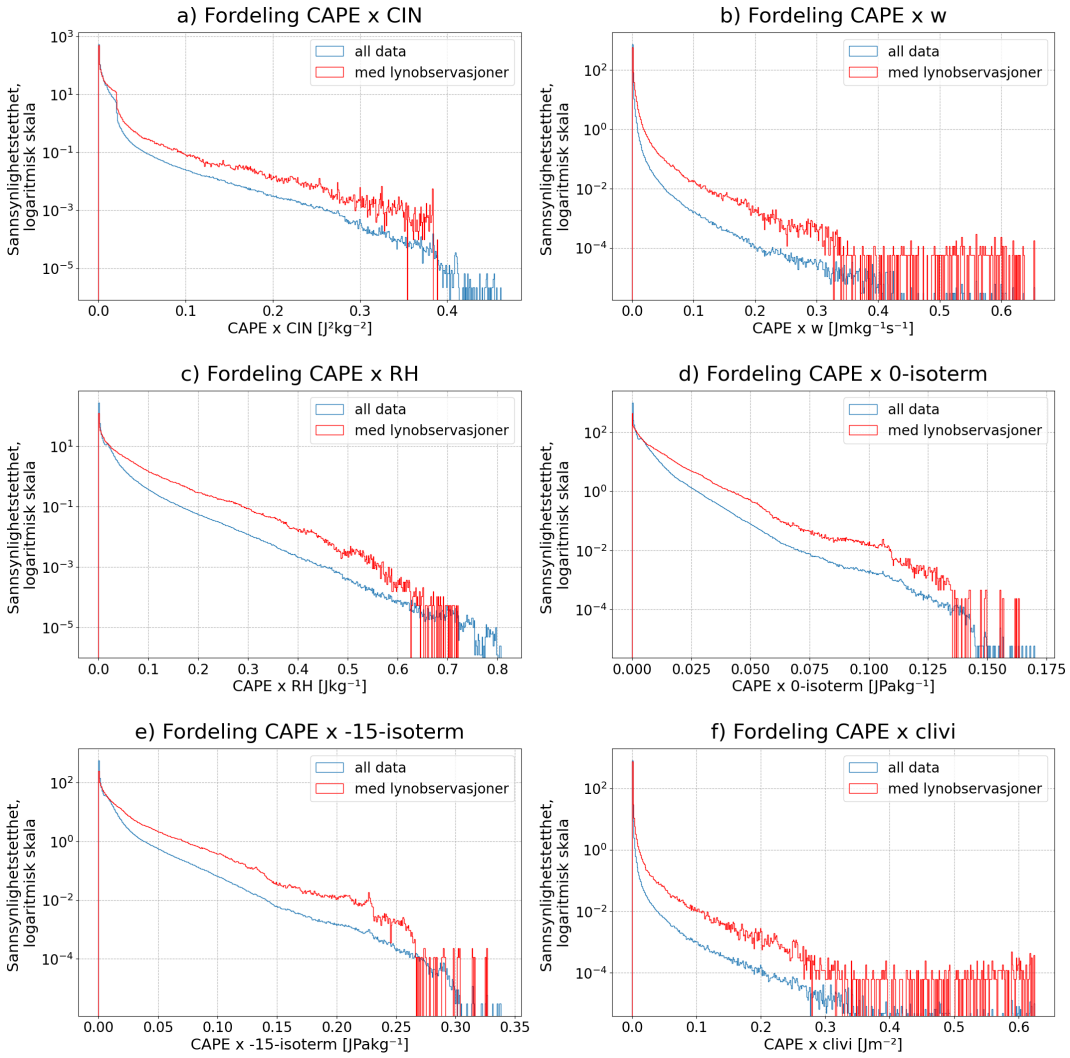
Preprosessering av parametere:
CAPE:
1) Beregne maksimal CAPE i hvert gitterpunkt over 6t periode
2) kjør gridpp med r=7 og quantile=0.9 på (1)
3) (Output fra (2))**(1/8)
CIN:
1) Beregne laveste CIN i hvert gitterpunkt over 6t periode
2) kjør gridpp med r=7 og quantile=0.1 på (1)
3) (output fra (2))**(1/4)
Vertikal hastighet:
1) Beregn maksimal W over alle trykknivåer for hvert gitterpunkt over en 6t periode
2) kjør gridpp med r=7 quantile=0.9
Relativ fuktighet 700hPa:
1) Beregne maksimal RH i 700 hPa i hvert gitterpunkt over en 6t periode
2) Kjør gridpp r=7 quantile=0.9

Deretter samles alle fire filene i ei fil med CDO, og lynindeksen beregnes med

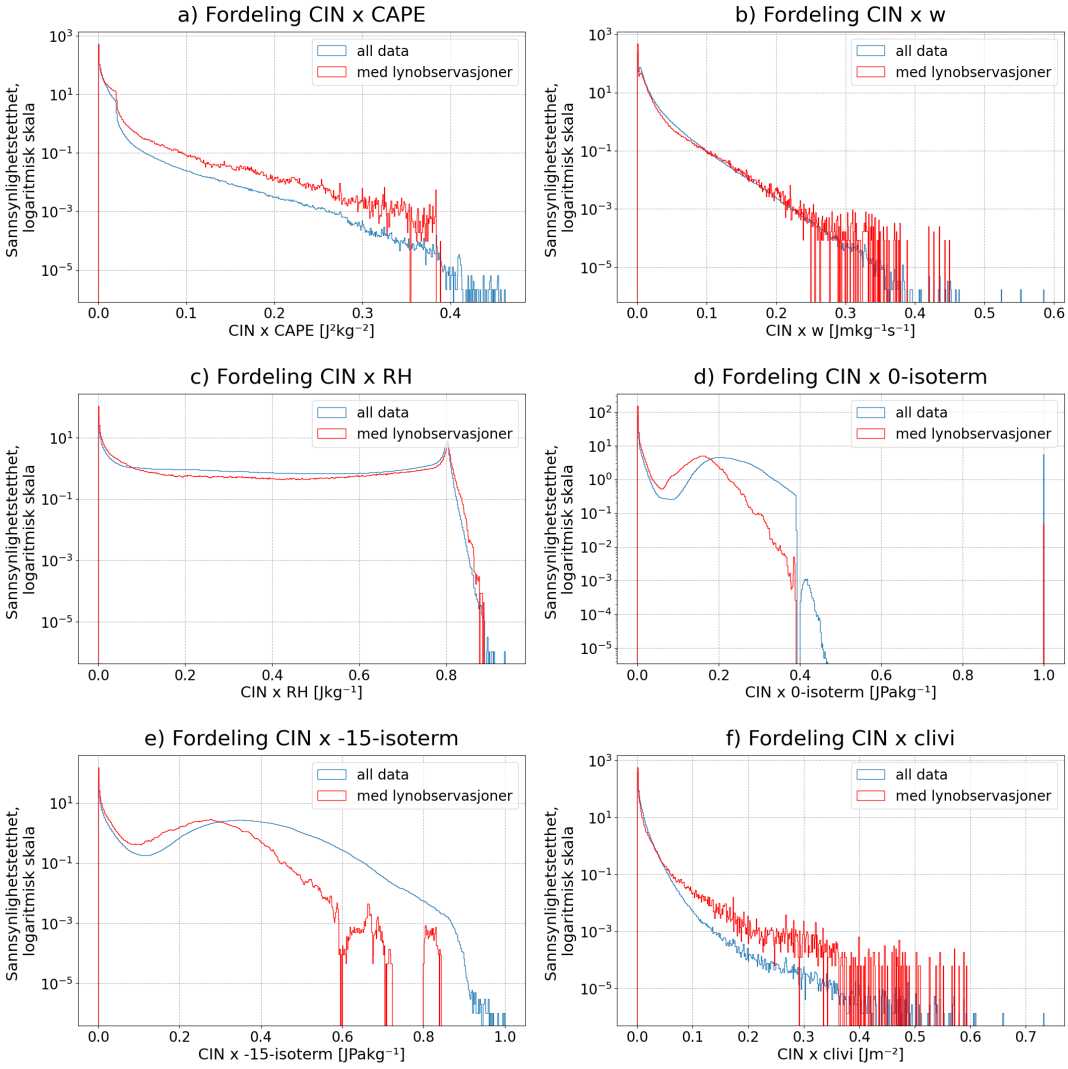
cdo -expr,'prob6hr=1/(1+exp(-(b0+b1*CAPE+ b2*CIN +c3*W + c4*RH)))' input.nc output.nc

b0 = -9.127936 b1 (koeffisienten foran CAPE) = 2.696786
b2 (koeffisienten foran CIN) = -0.001227567
b3 (koeffisienten foran W) = 0.6036084
b4 (koeffisienten foran RH) = 0.8730381

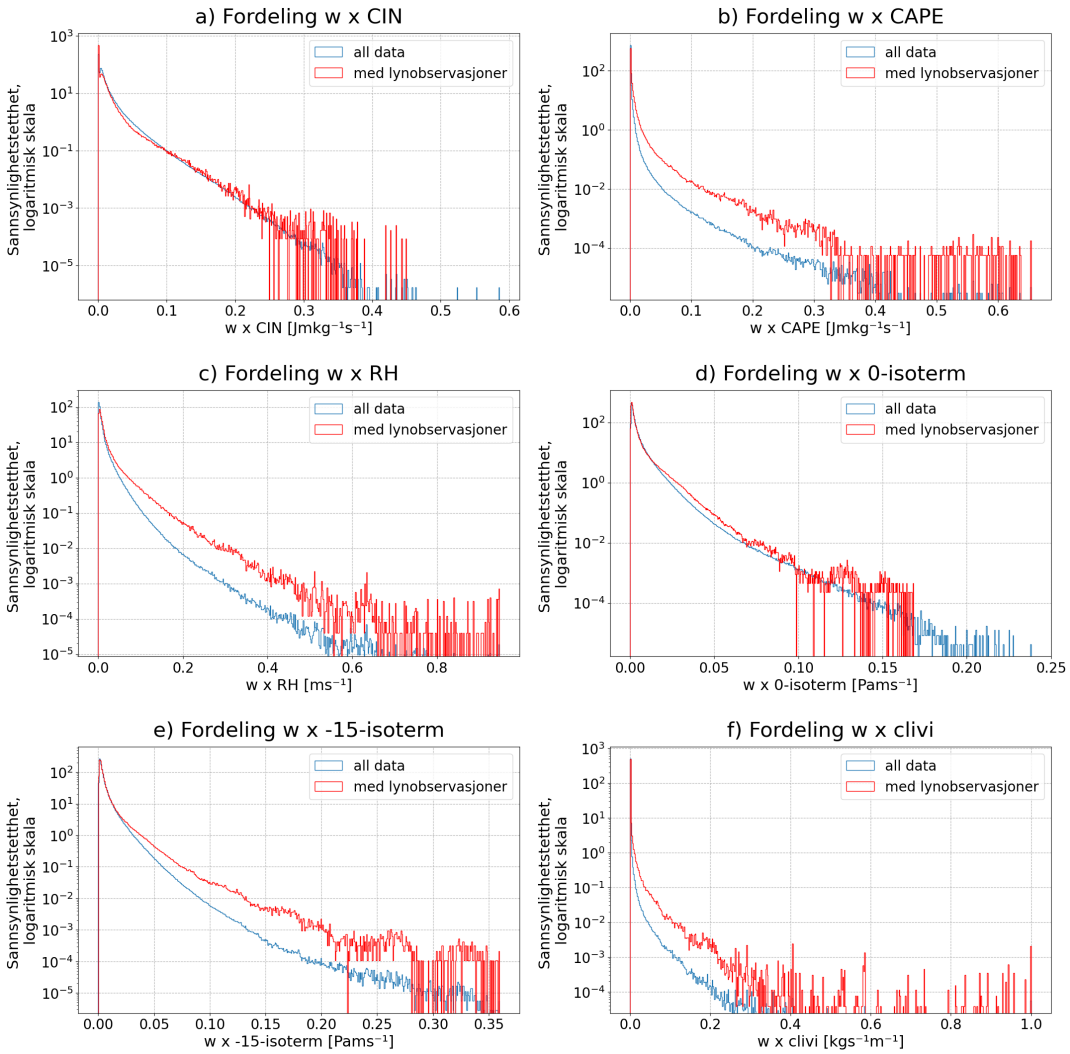
Vedlegg B. Multipliserte parametere



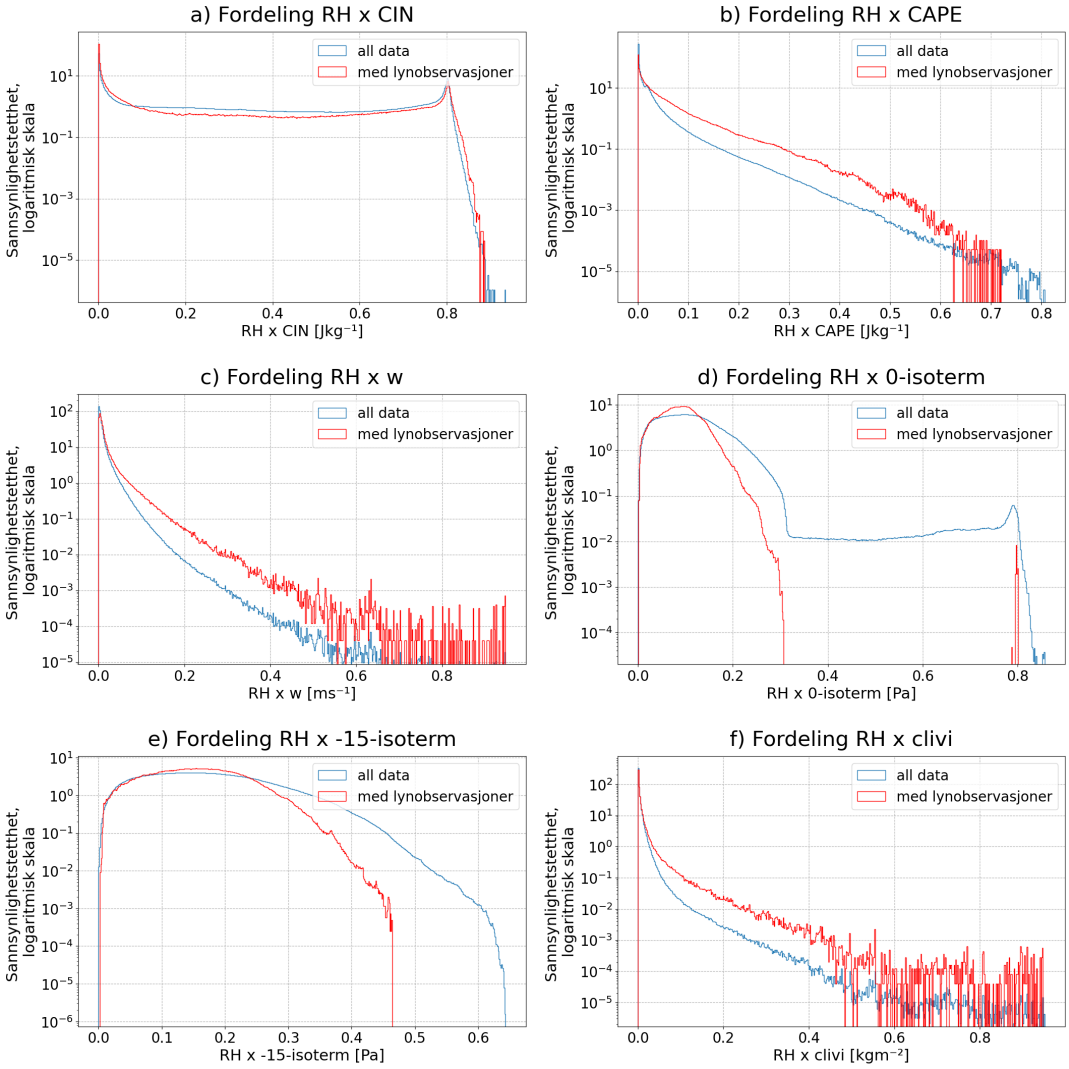
Figur B.1: Sannsynlighetstetthetsplot av fordelingen av alle parametere multiplisert med CAPE i en semi-logaritmisk skala. Blå farge er for hele datasettet, mens rød farge er for datapunkter som inneholder lynobservasjoner. Arealet under hver av grafene er lik 1.



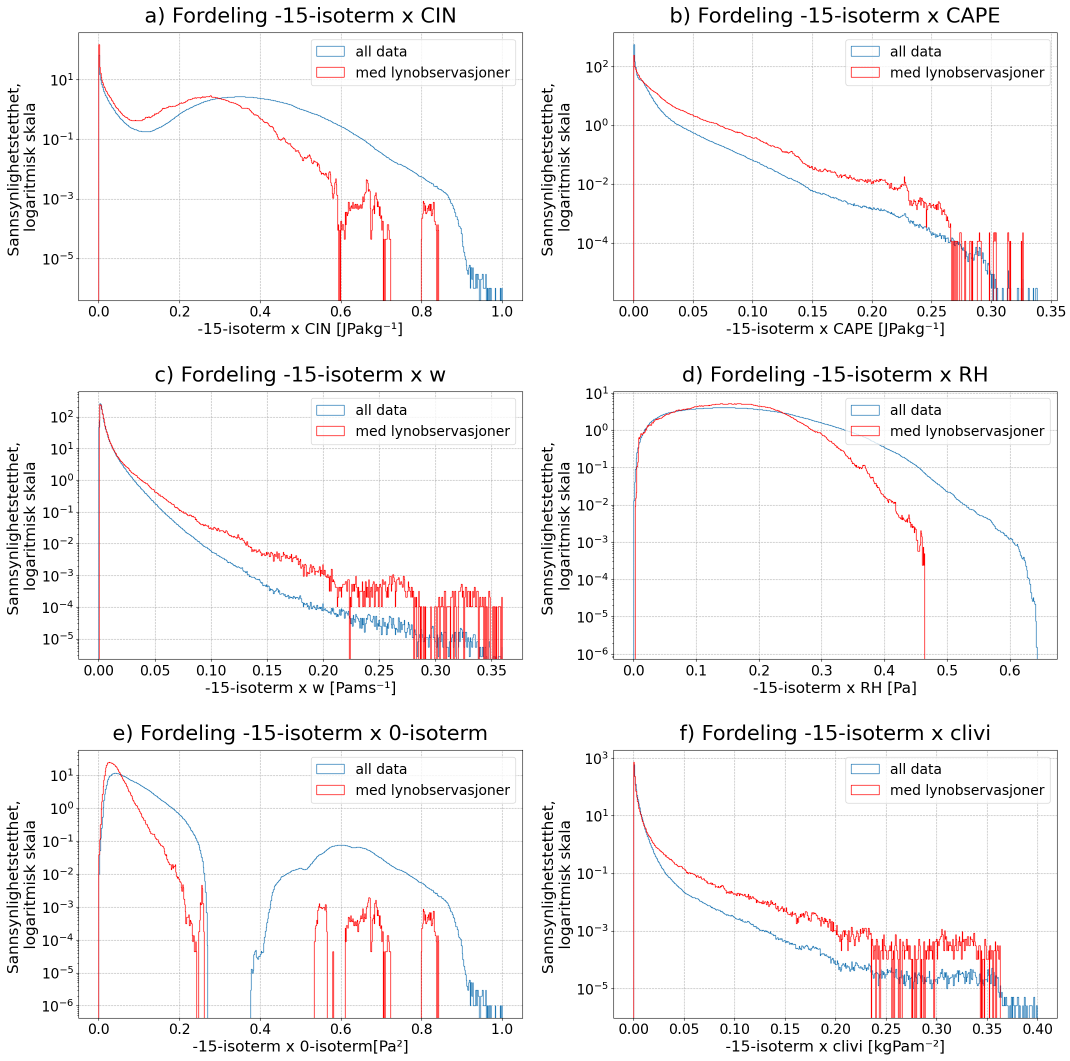
Figur B.2: Sannsynlighetstetthetsplot av fordelingen av alle parametere multiplisert med CIN i en semi-logaritmisk skala. Blå farge er for hele datasettet, mens rød farge er for datapunkter som inneholder lynobservasjoner. Arealet under hver av grafene er lik 1.



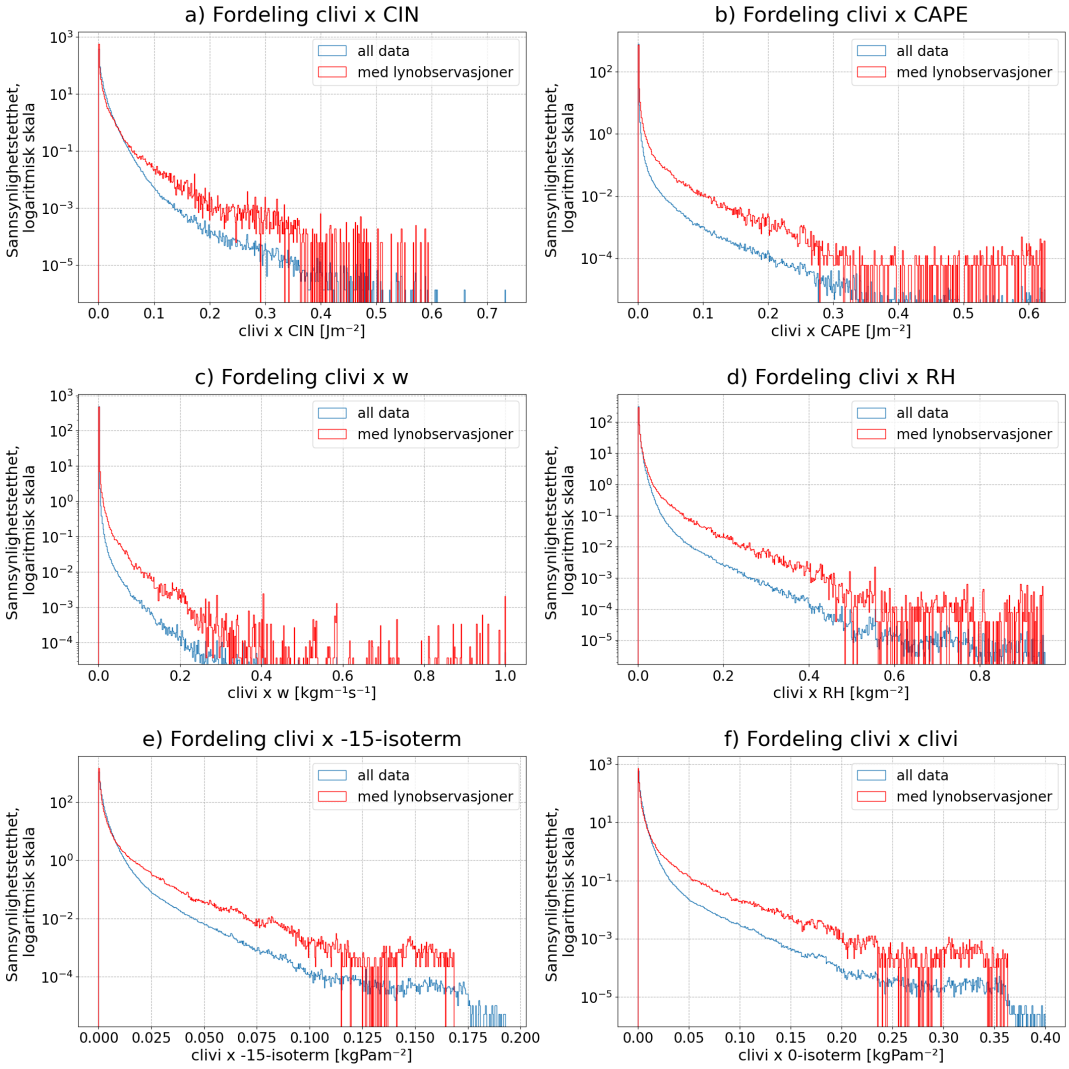
Figur B.3: Sannsynlighetstetthetsplot av fordelingen av alle parametere multiplisert med w i en semi-logaritmisk skala. Blå farge er for hele datasettet, mens rød farge er for datapunkter som inneholder lynobservasjoner. Arealet under hver av grafene er lik 1.



Figur B.4: Sannsynlighetstetthetsplot av fordelingen av alle parametere multiplisert med RH i en semi-logaritmisk skala. Blå farge er for hele datasettet, mens rød farge er for datapunkter som inneholder lynobservasjoner. Arealet under hver av grafene er lik 1.

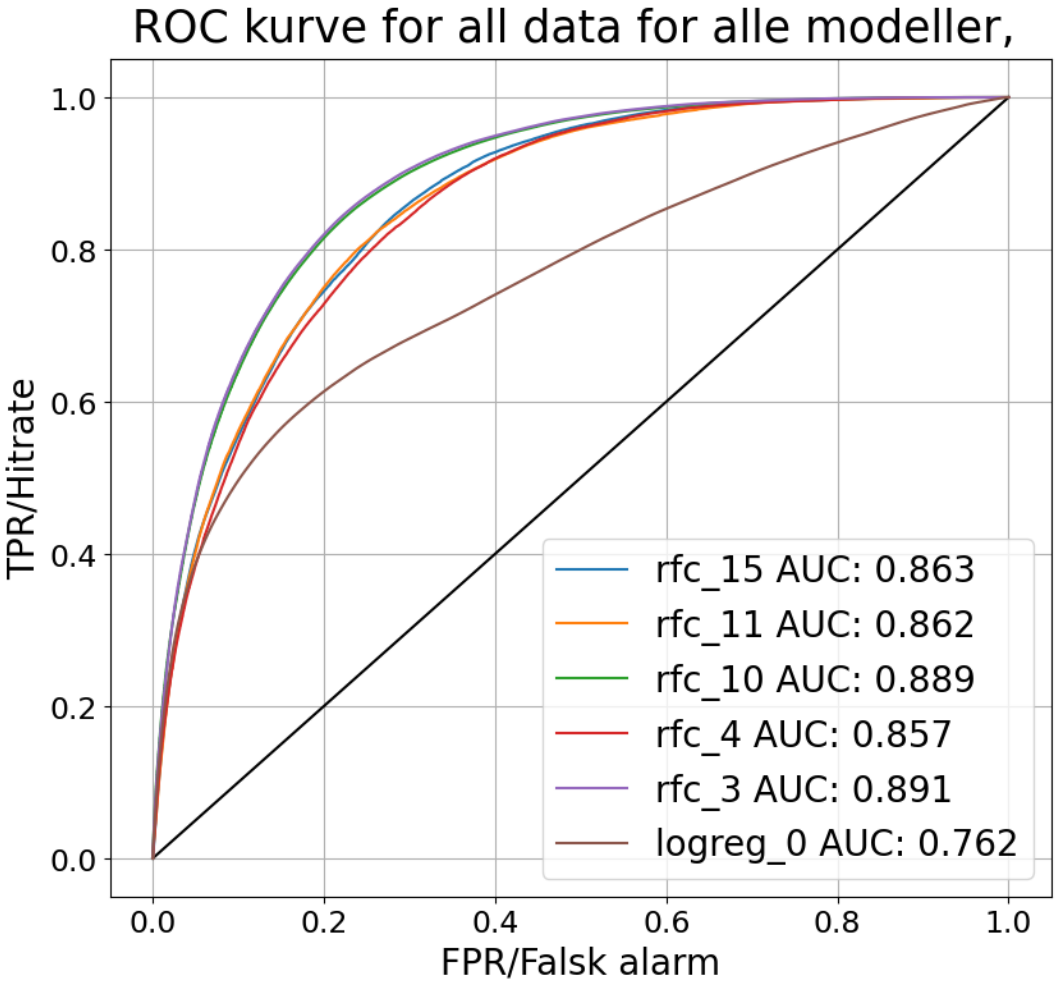


Figur B.5: Sannsynlighetstetthetsplot av fordelingen av alle parametere multiplisert med -15-isotermen i en semi-logaritmisk skala. Blå farge er for hele datasettet, mens rød farge er for datapunkter som inneholder lynobservasjoner. Arealet under hver av grafene er lik 1.



Figur B.6: Sannsynlighetstetthetsplot av fordelingen av alle parametere multiplisert med clivi i en semi-logaritmisk skala. Blå farge er for hele datasettet, mens rød farge er for datapunkter som inneholder lynobservasjoner. Arealet under hver av grafene er lik 1.

Vedlegg C. ROC-kurve all data 2014-2018



Figur C.1: ROC-kurver for all treningsdata for diverse modeller trent på all data fra 2014-2018.

Norges miljø- og
biovitenskapelige
universitet