

DANMARKS METEOROLOGISKE INSTITUT
TEKNISK RAPPORT

02-28

Logistisk Kalman filter for kraftig nedbør

December 2002

Michael Steffensen

ISSN 0906-897X
ISSN 1399-1388 (online)



København 2002

Indholdsfortegnelse

1.	Indledning.....	2
2.	Data.....	3
3.	Kategoriske prognoser for 24 timers akkumuleret nedbør.....	4
4.	Sandsynlighedsprognoser.....	8
5.	Logistisk regression og Kalman filter	10
6.	Verifikation af sandsynlighedsprognoser.....	14
6.1	Pålidelighed og skarphed.....	14
6.2	Brier score og skill score	15
6.3	Cost/Loss Value	16
6.3.1	Kategoriske prognoser	17
6.3.2	Sandsynlighedsprognoser.....	18
6.4	ROC kurver	20
7.	Konklusion	21

1. Indledning

Denne rapport beskriver nogle mulige metoder til automatiske forudsigelse af nedbør og kraftig/meget nedbør ved hjælp af DMI-HIRLAM data og synop observationer.

Forudsigelse af især kraftig/meget nedbør har vist sig at være vanskelig. Derfor fokuseres der på at bestemme sandsynligheder for nedbør større end en given tærskelværdi i modsætning til kategoriske prognoser for nedbør.

Fordelen ved sandsynlighedsprognoser er, at de samtidig med den aktuelle prognose giver en vurdering af prognosens usikkerhed, hvilket mange brugere kan drage nytte af. I den sammenhæng anvendes ofte et relativt nyt mål for verifikation, det såkaldte Cost/Loss Value, som viser hvilken økonomisk værdi en prognose har for en given bruger beskrevet ved forholdet mellem brugerens udgift til at beskytte sin aktivitet mod, i dette tilfælde kraftig nedbør, og tabet ved mulige konsekvenser af kraftig nedbør uden at have beskyttet sig. Beskyttelse skal i denne sammenhæng forstås meget bredt og kan for eksempel bestå i at undlade at udføre en aktivitet med deraf følgende konsekvenser, for eksempel produktionstab, der så repræsenterer udgiften til beskyttelse.

En af de mulige metoder til bestemmelse af sandsynligheder for nedbør er logistisk regression med et logistisk Kalman filter. Et logistisk Kalman filter er en pendant til et almindeligt Kalman filter, som adaptivt justerer koefficienterne i en lineær regression mellem prognoseværdier og observerede værdier. Logistisk regression mellem prognoseværdier og observerede værdier bestemmer koefficienter i en algoritme, som giver sandsynligheder for nedbørsmængder større end den givne tærskelværdi givet en aktuel (kategorisk) prognoseværdi for nedbørsmængden. Et logistisk Kalman filter justerer adaptivt koefficienter i regressionen.

Kapitel 2 beskriver kort de anvendte data. I kapitel 3 karakteriseres DMI-HIRLAM's kategoriske prognose for nedbør ved sammenligning med observationer fra en synop station. Herefter bestemmes i kapitel 4 sandsynligheder for nedbørsmængder større end en given tærskelværdi ud fra en simpel tællemetode. I kapitel 5 vises, hvorledes disse sandsynligheder også kan bestemmes med logistisk regression og hvorledes et logistisk Kalman filter defineres. Kapitel 6 gennemgår forskellige metoder til verifikation af sandsynlighedsprognoser.

2. Data

Observeret 24 timers akkumuleret nedbør, $ObsN(dato_o)$, er beregnet ud fra synop observationer klokken 6, 12 og 18 UTC ved:

$$ObsN(dato_o) = N18(dato_o-1) - N12(dato_o-1) + N06(dato_o) + N12(dato_o)$$

Hvor $dato_o$ er observations dato for 24 timers akkumuleret nedbør fra klokken 12 forrige dag ($dato_o-1$) til klokken 12 $dato_o$. $N6(dato_o)$ og $N18(dato_o-1)$ er observeret 12 timers akkumuleret nedbør henholdsvis klokken 6 $dato_o$ og klokken 18 forrige dag, $dato_o-1$.

$N12(dato_o)$ og $N12(dato_o-1)$ er observeret akkumuleret 6 timers nedbør klokken 12 $dato_o$ og klokken 12 forrige dag, $dato_o-1$, se illustration nedenfor.

```

06 dato_o-1      12 dato_o-1  18 dato_o-1  00 dato_o      06 dato_o      12 dato_o
I-----I+++++I+++++I+++++I+++++I+++++I
I-N12(dato_o-1)--I
I-----N18(dato_o-1)-----I-----N6(dato_o)-----I--N12(dato_o)--I

```

Prognosen for nedbør er DMI-HIRLAM parametrene ACPR (konvektiv) og ASPR (stratiform) akkumuleret nedbør fra analyse tidspunktet til prognose tidspunktet. Der er lavet prognoseværdier for akkumuleret nedbør for dels 00 til +24 timer, $N(00 \text{ til } 24)$, og dels +24 til +48 timer $N(24 \text{ til } 48)$ ud fra klokken 12 UTC analyser.

$$N(00 \text{ til } 24)(dato_a) = ACPR(+24)(dato_a) + ASPR(+24)(dato_a)$$

er den akkumulerede nedbør fra analysetidspunktet klokken 12UTC $dato_a$ og 24 timer frem.

$$N(24 \text{ til } 48)(dato_a) = ACPR(+48)(dato_a) + ASPR(+48)(dato_a) - [ACPR(+24)(dato_a) + ASPR(+24)(dato_a)]$$

er den akkumuleret nedbør fra analysetidspunktet klokken 12UTC $dato_a$ plus 24 timer og yderligere 24 timer frem.

Samhørende værdier af observation og prognose $N(00 \text{ til } 24)(dato_a)$ med analysetidspunkt $dato_a$ fås ved at bruge observationen med dato $dato_o = dato_a+1$ så:

$$N(00 \text{ til } 24)(dato_a) \langle \rangle ObsN(dato_a+1)$$

Samhørende værdier af observation og prognosen $N(24 \text{ til } 48)(dato_a)$ med analysetidspunkt $dato_a$ fås ved at bruge observationen med dato $dato_o = dato_a+2$ så:

$$N(24 \text{ til } 48)(dato_a) \langle \rangle ObsN(dato_a+2)$$

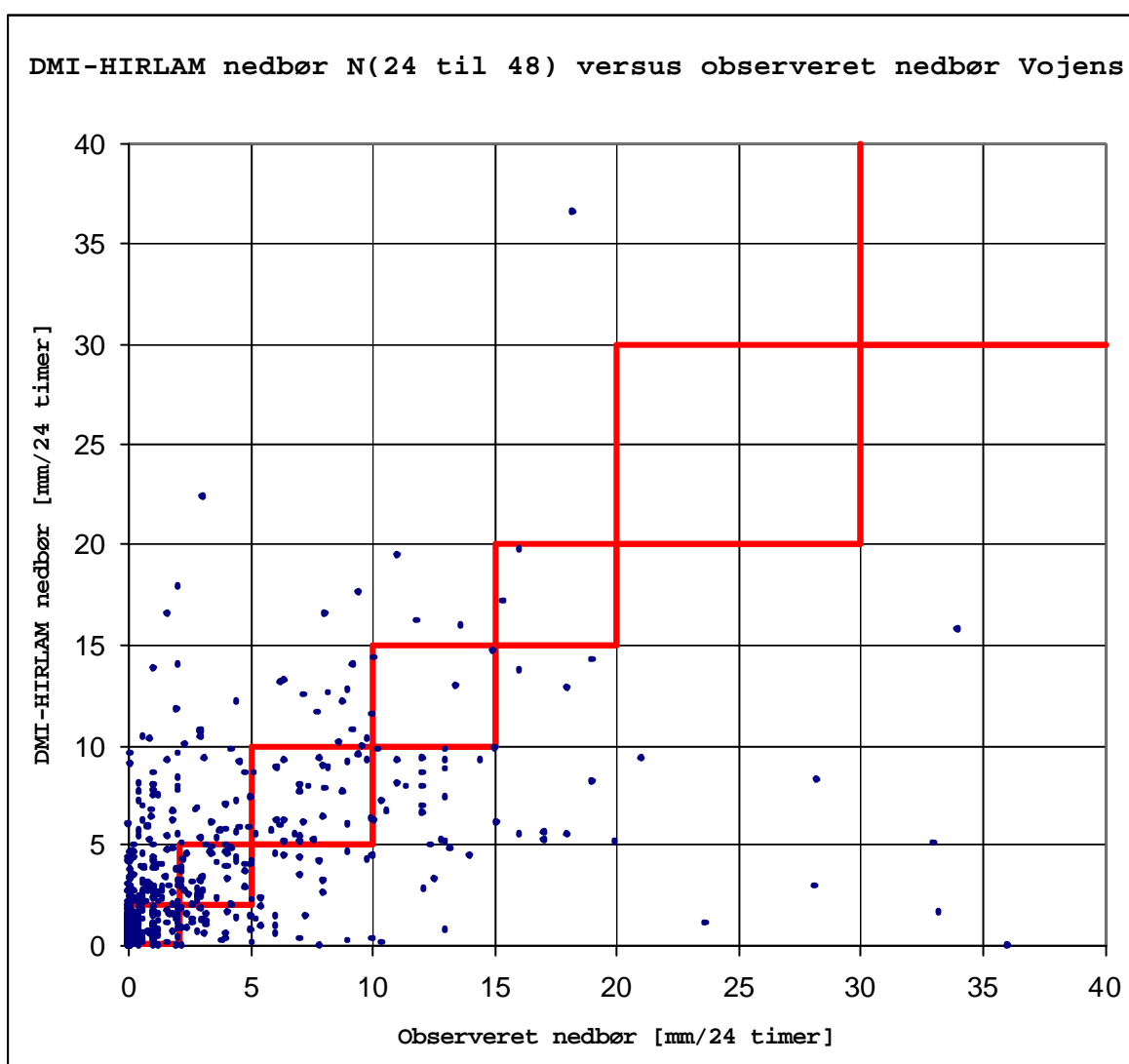
Der er lavet statistik for tærskelværdierne >2 , >5 , >10 , >15 og >20 [mm/24 timer], hvor nedbør >10 , >15 og >20 [mm/24 timer] kan betragtes som kraftig nedbør.

Data dækker perioden 9/9 1999 til 10/12 2002 . Dog mangler data for nogle af dagene. Nedbørsobservationer fra Vojens 6110 er blevet brugt i verifikationen.

3. Kategoriske prognoser for 24 timers akkumuleret nedbør

Scatterplot af 24 timers akkumuleret nedbør fra DMI-HIRLAM versus observeret 24 timers akkumuleret nedbør fra tættest liggende SYNOP station giver et godt overblik over nedbørsprognosers evne til at forudsige nedbørsmængden. Figur 1 viser et scatterplot for prognoser af akkumuleret nedbør fra +24 til +48 timer i forhold til analyse tidspunktet for prognosen. Punkterne ligger relativt spredt og illustrerer dermed vanskeligheden ved at forudsige nedbørsmængder.

Prognoserne kan også beskrives med numeriske verifikationsmål, hvilket gør det lettere umiddelbart at sammenligne to forskellige prognoser. Man skal dog være opmærksom på, at sådanne numeriske beskrivelser kun beskriver et bestemt aspekt ved prognosen.



Figur 1 Scatterplot DMI-HIRLAM nedbør versus observeret nedbør ved Vojens synop station.

Dertil kommer, at de numeriske prognoser beregner areal middel nedbørsmængder i modsætning til synop målestationer, som er punkt målinger, hvilket gør det vanskeligt at lave en

direkte sammenligning. For eksempel er der mulighed for, at en kraftig byge i området ikke rammer nedbørsmåleren. Dette udgør en vanskelighed ved verifikation af nedbørsmængder.

Root Mean Square Error, RMSE, for prognosen er 5.5 [mm/24timer]. Middel differensen på kun -0.67 [mm/24timer] viser, at prognosen kun overestimerer nedbørsmængden en lille smule. Det betyder, at der ikke vil være nogen gevinst ved at lave en lineær regression af nedbør, eventuelt med et Kalman filter, som har vist anvendelig for andre meteorologiske parametre, fordi lineær regression primært kan reducere en eventuel bias.

De røde kvadrater i figur 1 viser intervallerne i en kontingenstabel. Kontingenstabellen viser i princippet det samme som et scatterplot, men med den forskel, at værdier for henholdsvis prognoser og observationer samles i nogle valgte intervaller. De anvendte intervaller er [0-0.1], [0.1-2], [2-5], [5-10], [10-15], [15-20], [20-30] og >30 [mm/24timer].

Tabel 1 viser kontingenstabellen svarende til scatterplottet i figur 1. Diagonal elementerne i matricen er de hændelser, hvor prognose og observation falder i samme interval og betegnes derfor som et hit (korrekte). Værdier over diagonalen repræsenterer hændelser, hvor prognosen underestimerer nedbørsmængden, og værdier under diagonalen repræsenterer hændelser, hvor prognosen overestimerer nedbørsmængden. Hit rate er summen af alle korrekte prognoser, altså summen af værdierne i diagonal elementerne, relativt til det totale antal prognoser i data sættet. Hit rate for kontingenstabellen er 35 %.

Ved små observerede nedbørsmængder altså intervallerne [0-0.1] og [0.1-2] [mm/24timer] er der ganske mange hændelser, hvor prognosen ligger i intervaller ovenover [0.1-2], [2-5], [5-10] og [10-15]. Disse hændelser under diagonalen i tabellen viser, at prognosen overestimerer nedbørsmængden ved små nedbørsmængder. Omvendt ses, at værdierne over diagonalen viser, at prognosen underestimerer nedbørsmængden ved større nedbørsmængder.

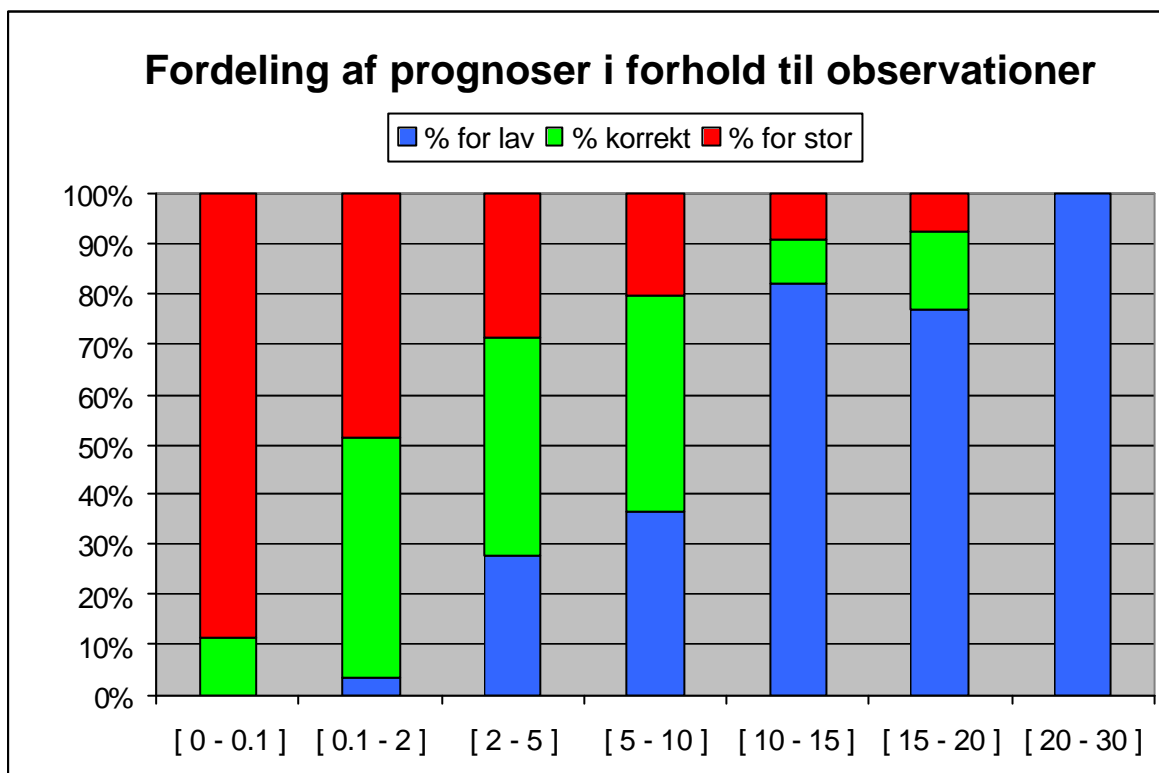
Prognose	Observeret								
	[0 - 0.1]	[0.1 - 2]	[2 - 5]	[5 - 10]	[10 - 15]	[15 - 20]	[20 - 30]	[30 - 60]	
[0 - 0.1]	98	6	1	1	0	0	0	1	107
[0.1 - 2]	196	82	27	11	2	0	1	1	320
[2 - 5]	32	54	44	13	4	0	1	1	149
[5 - 10]	3	23	23	29	21	7	2	2	110
[10 - 15]	0	5	5	12	3	3	0	0	28
[15 - 20]	0	2	0	2	3	2	0	1	10
[20 - 30]	0	0	1	0	0	0	0	0	1
[30 - 60]	0	0	0	0	0	1	0	0	1
	329	172	101	68	33	13	4	6	

Tabel 1 kontingenstabel.

Tallene under matricen og i kolonnen helt til højre viser hvor mange henholdsvis observationer og prognoser, der er i hvert interval. Disse tal beskriver klimatologien af henholdsvis de observerede nedbørsmængder og prognoserne. Her ses tydeligt prognosens overestimering af små nedbørsmængder, idet prognosen kun har 107 hændelser i intervallet [0-0.1], mens der er hele 329 observerede hændelser i samme interval og omvendt 320 hændelser i prognosens interval [0.1-2], mens der kun er 172 observerede hændelser i dette interval.

Figur 2 viser fordelingen af prognoser indenfor intervallerne i forhold til observationsintervallet. Røde søjler er andelen af prognoser i intervaller med større nedbørsmængder end observationsintervallet, grønne søjler er andelen af prognoser i samme interval som observationsintervallet (korrekte) og blå søjler er andelen af prognoser i intervaller med mindre nedbørsmængder end observationsintervallet.

De relativt store grønne søjler for intervallerne [0.1-2], [2-5] og [5-10] viser, at prognosen er bedst i disse intervaller, men også for disse intervaller er der en del usikkerhed med mange prognoser i intervaller med både større og mindre nedbørsmængder end observationen.



Figur 2 Fordelingen af prognoser indenfor intervallerne i forhold observationsintervallet.

De store røde og blå søjler viser hvordan prognosen overestimerer ved små nedbørsmængder og underestimerer ved større. Det er altså relativt mere vanskeligt at forudsige små og store nedbørsmængder. For store nedbørsmængder er der imidlertid den vanskelighed, at der er få hændelser til at få en ordentligt statistik. Der er kun 56 hændelser ud af 726 med observerede nedbørsmængder større end 10 [mm/24 timer].

En metode til at håndtere forudsigelse af store nedbørsmængder er at lave udsigter for nedbørsmængder større end en givne tærskelværdier, $N_{\text{tærskel}}$, for eksempel 10, 15 eller 20 [mm/24timer]. Det kan gøres ved at forudsige nedbørsmængden $> N_{\text{tærskel}}$, når prognosens værdi for nedbørsmængden er $> N_o$, hvor N_o skal bestemmes så forudsigelsen bliver optimal. N_o er således typisk forskellig fra $N_{\text{tærskel}}$. For store nedbørsmængder gælder at $N_o > N_{\text{tærskel}}$.

Kontingenstabeller for sådanne prognoser bliver 2*2 matricer svarende til de to hændelser; nedbørsmængder $> N_{\text{tærskel}}$ og nedbørsmængder $\leq N_{\text{tærskel}}$ for henholdsvis prognose og ob-

ervation. En måde at få optimale prognoser på er for eksempel at bestemme N_o så hit rate for kontingenstabellen bliver størst.

Tabel 2 a) og b) viser 2*2 kontingenstabeller for nedbørsmængde > 10 [mm/24timer]. Tabel 2a) er baseret på forudsigelser af nedbørsmængder større end 10 [mm/24 timer], når prognosen er større end 10 [mm/24 timer], altså $N_o = 10$ [mm/24timer] og tabel 2b) er baseret på forudsigelser af nedbørsmængder større end 10 [mm/24 timer], når prognosen er større end 14 [mm/24 timer], altså $N_o = 14$ [mm/24timer].

Prognose	Observeret	
	nedbør>10	nedbør<10
nedbør>10	2	6
nedbør<10	24	294

Tabel 2a)

Prognose	Observeret	
	nedbør>10	nedbør<10
nedbør>10	2	0
nedbør<10	24	300

Tabel 2b)

Hit rate for tabel 2a) er $(2+294)/326 = 91\%$, mens hit rate for tabel 2b) er $(2+300)/326 = 93\%$. Imidlertid ses umiddelbart af matricerne, at hit rate ikke beskriver kvaliteten af prognosen helt, idet antallet af ikke forudsagte hændelser med nedbørsmængde større end 10 [mm/24 timer] er hele 24, hvilket skal ses i sammenhold med antallet af gange denne hændelse er forudsagt korrekt, som kun er 2 gange.

Alternativet til kategoriske prognoser er sandsynlighedsprognoser, som samtidig med den aktuelle prognose giver en vurdering af prognosens usikkerhed. Disse vil blive beskrevet i næste kapitel.

4. Sandsynlighedsprognoser

I dette projekt er sandsynligheder for nedbør større end tærsklerne: 2, 5, 10, 15, og 20 [mm/24timer] blevet analyseret. I rapporten vil resultater for nedbørsmængder større end 10 [mm/24timer] blive anvendt som eksempel.

Sandsynlighedsprognoser tager udgangspunkt i den netop beskrevet kategoriske prognose for nedbørsmængder større end en given tærskel værdi, ved i stedet at forudsige en sandsynlighed for nedbørsmængder større end en given tærskel værdi ud fra prognosen af nedbørsmængden

Kontingenstabellen i tabel 1 kan bruges til at vurdere størrelsen af disse sandsynligheder ud fra hyppigheder ved en simpel optælling. Hvis prognose værdien for eksempel ligger i intervallet [5-10] kan sandsynligheden for nedbør større end 10 [mm/24timer] vurderes som af summen af værdier i kontingenstabellens række 4 (intervallet [5-10]) fra og med søjle 5 (intervallet [10-15]) til og med søjle 8 (intervallet [30-60]), det vil sige summen over alle intervaller med observeret nedbørsmængde > 10 [mm/24timer], divideret med det totale antal prognoser i intervallet [5-10], som findes i kolonnen yderst til højre. Ved at bruge værdier fra tabel 1 fås altså $21+7+2+2=32$ divideret med 110, hvilket giver en sandsynlighed på 0.29 for nedbørsmængder større end 10 [mm/24 timer], når den kategoriske prognose for nedbørsmængden ligger i intervallet [5-10].

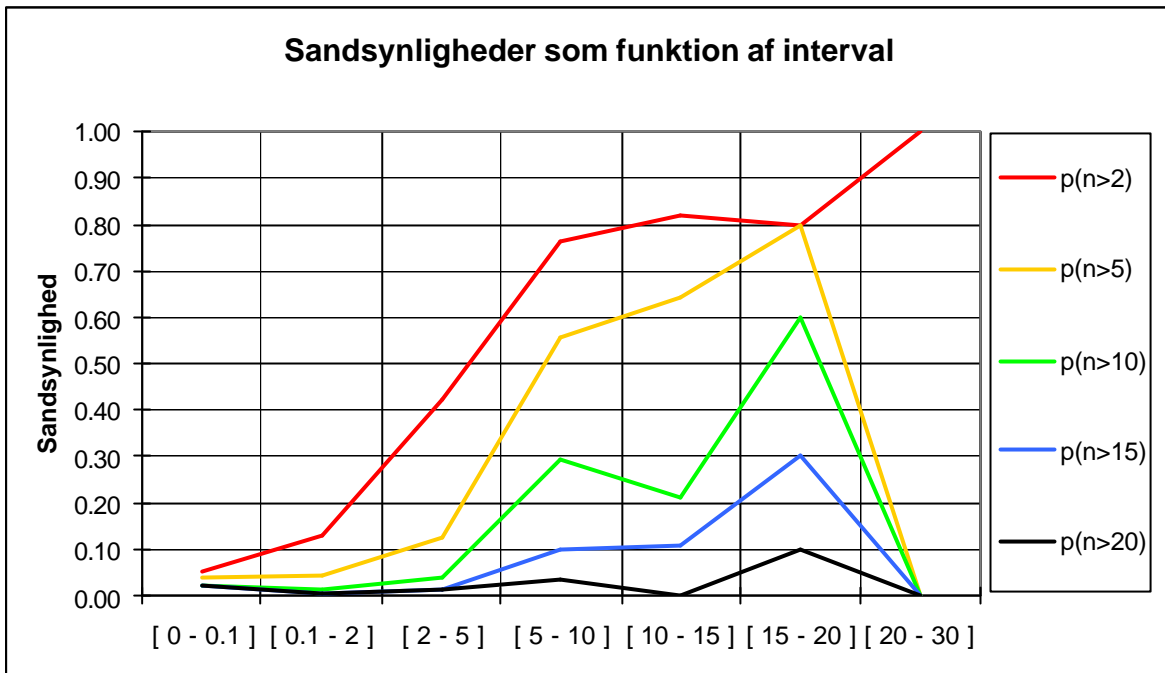
På tilsvarende vis vurderes sandsynligheden for nedbørsmængder større end 10 [mm/24timer], når prognosen ligger i intervallet [10-15], som af summen af værdier i kontingenstabellens række 5 (intervallet [10-15]) fra og med igen søjle 5 ([5-10]) til søjle 8 (intervallet [30-60]), divideret med det totale antal prognoser i intervallet [10-15], altså $3+3=6$ divideret med 28, hvilket giver 0.21.

Sådan kan sandsynligheden for nedbørsmængder større end 10 [mm/24timer] vurderes for prognose værdier i alle intervallerne [0.1-2], [2-5], [5-10], [10-15], [15-20], [20-30] [mm/24timer]. På tilsvarende måde kan sandsynlighederne for de øvrige tærskelværdier 2, 5, 15 og 20 bestemmes ud fra kontingenstabellen i tabel 1. Disse sandsynligheder er vist i tabel 3. For eksempel er sandsynligheden for nedbørsmængde større end 5 [mm/24 timer] når den kategoriske prognose ligger i intervallet [10-15] 0.64.

	p(n>2)	p(n>5)	p(n>10)	p(n>15)	p(n>20)
[0 - 0.1]	0.06	0.04	0.03	0.03	0.03
[0.1 - 2]	0.13	0.05	0.01	0.01	0.01
[2 - 5]	0.42	0.13	0.04	0.01	0.01
[5 - 10]	0.76	0.55	0.29	0.10	0.04
[10 - 15]	0.82	0.64	0.21	0.11	0.00
[15 - 20]	0.80	0.80	0.60	0.30	0.10
[20 - 30]	1.00	0.00	0.00	0.00	0.00

Tabel 3 Sandsynlighederne for nedbør større end tærskelværdier 2, 5, 10, 15 og 20 [mm/24 timer]

Figur 3 viser kurverne for sandsynligheden for nedbørsmængder større end tærskelværdierne 2, 5, 10, 15 og 20 [mm/24 timer] som funktion af intervallerne i kontingenstabellen.



Figur 3 Sandsynligheden for nedbørsmængder >tærskelværdierne 2, 5, 10, 15 og 20 [mm/24 timer]

For en given tærskelværdi vil man forvente, at sandsynligheden stiger, når man går til et højere interval. Dette er ikke helt tilfældet i den viste tabel og figur. Det skyldes, at nedbørsmængder større end 10, 15 og 20 forekommer relativt sjældent, så data materialet ikke er stort nok til at få tilstrækkeligt mange hændelser med. Desuden vil kurvernes aktuelle forløb også afhænge af de valgte intervalinddelinger, hvilket vil have størst betydning for de store nedbørsmængder med få hændelser. Hvis der kun er to hændelser i et valgt interval, hvor den ene er korrekt og den anden er falsk, kan man ændre intervalinddelingen så kun den korrekte findes i intervallet, hvorved hyppigheden af korrekte hændelser i intervallet ændres fra 50 til 100 %

I næste kapitel vises, hvordan dette til dels kan afhjælpes ved anvendelse af logistisk regression, som giver sandsynligheden for nedbørsmængder større end en tærskelværdi som funktion af enhver prognoseværdi uden brug af valgte intervaller.

5. Logistisk regression og Kalman filter

Vurderingen af sandsynligheder for nedbørsmængder større end tærsklerne 2, 5, 10, 15, 20 [mm/24 timer], som blev illustreret i forrige kapitel, kan beregnes for alle prognose værdier uafhængigt af de valgte intervaller ved hjælp af logistisk regression. Med logistisk regression kan man for enhver prognose værdi beregne sandsynligheder for nedbørsmængde større end tærskelværdien. Denne metode kan anvendes ved udarbejdelse af automatiske prognoser.

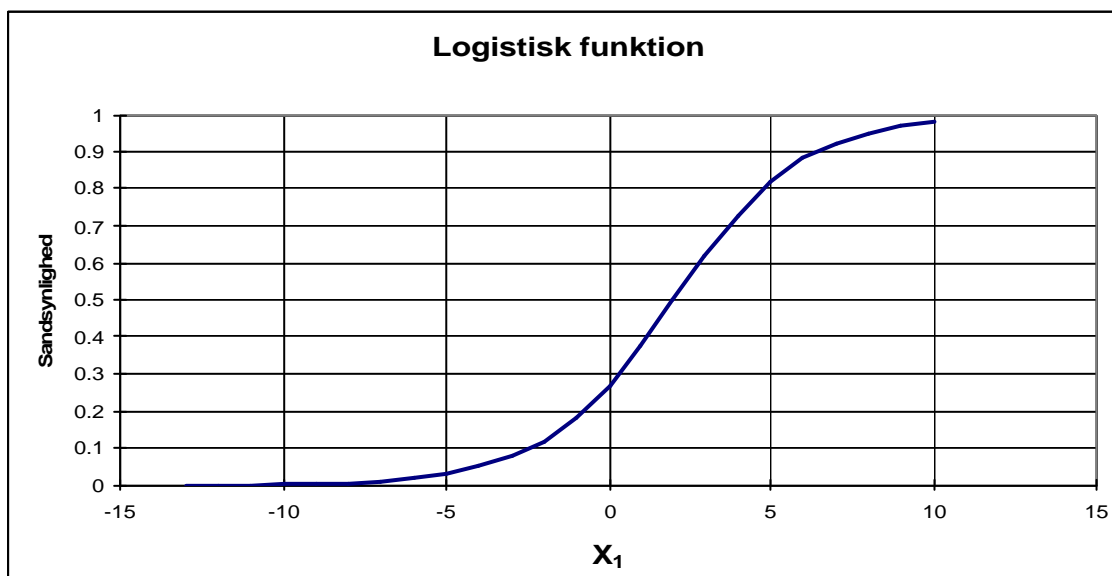
Logistisk regression er mægt til normal regression, men med den forskel at den afhængige variabel, Y , kun kan antage heltalsværdier (eller diskrete værdier), i dette tilfælde 1 hvis nedbørsmængden større end tærskelværdien og 0 ellers.

I normal regression af en afhængig variabel Y ud fra et sæt af uafhængige variable X_i , $i=1, \dots, N$ haves ligningen $\hat{Y} = a_0 + a_1 X_1 + \dots + a_N X_N$ for den estimerede værdi Y . Koefficienterne, a_i , bestemmes ud fra minimalisering af fejlfunktionen, som er summen af $(\hat{Y} - Y)^2$ over sættet af kendte Y 'er og estimaterne \hat{Y} .

I logistisk regression transformeres ligningen $\hat{Y} = a_0 + a_1 X_1 + \dots + a_N X_N$ til en sandsynlighed ved hjælp af en logistisk funktion $P = P(a_0 + a_1 X_1 + \dots + a_N X_N)$, som antager værdier mellem 0 og 1. Den logistiske funktion er defineret ved:

$$P = \frac{e^{(a_0 + a_1 X_1 + \dots + a_N X_N)}}{1 + e^{(a_0 + a_1 X_1 + \dots + a_N X_N)}}$$

Forløbet af en logistisk funktion med én uafhængig variable X_1 er vist figur 4

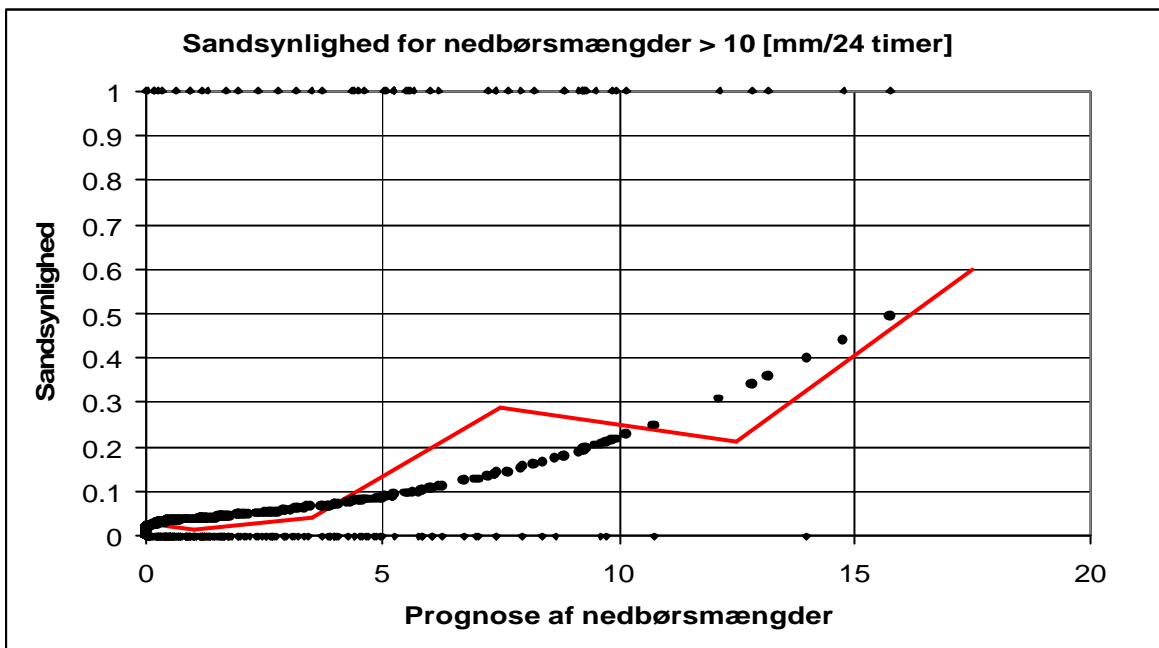


Figur 4 Forløbet af en logistisk funktion med én uafhængig variable X_1

I stedet for fejlfunktionen $(\hat{Y} - Y)^2$ benyttes i logistisk regression en fejlsandsynlighed bestemt ved $(1-P)^2$, hvor P er den estimerede sandsynlighed ud fra den logistiske funktion

$P(\hat{Y})$ og I er indikatorfunktionen, som er 1 hvis hændelsen indtræffer og 0 ellers. Koefficienterne, a_i bestemmes ved at summen af $(I-P)^2$ over data sættet skal være mindst mulig.

Figur 5 viser sandsynligheden for nedbørsmængder større end 10 [mm/24 timer], som funktion af den kategoriske prognose for nedbørsmængder. Den røde kurve viser sandsynligheden vurderet ud fra intervaller, hvor intervallet er blevet repræsenteret ved middelværdien af endepunkternes værdier. Der er god overensstemmelse mellem de to kurver. De sorte ruder markerer indikatorfunktions værdi. Der er end del markeringer for prognose værdier mindre end 10 [mm/24 timer].



Figur 5 Sandsynligheden for nedbørsmængder større end 10 [mm/24 timer], som funktion af den kategoriske prognose for nedbørsmængder

Kurveforløbet ligner ikke helt kurven i figur 4, når nedbørsmængderne er små. Det skyldes at den kategoriske prognose for nedbørsmængder kun kan være positiv eller nul. Fordelingen af nedbørsmængder fra kategoriske prognoser er derfor ikke normal fordelt, hvilket i dette tilfælde ville resultere i for store sandsynligheder, når den kategoriske prognose giver en lille nedbørsmængde. Derfor transformeres prognosen for nedbørsmængder, X , til $X' = \ln(X+tiny)$ hvis $X < 1$ og $X' = X-1$ når $X \geq 1$. Den transformerede prognose X' bliver kontinuert med en kontinuert første afledet. Tiny indføres fordi $\ln(0)$ ikke er defineret.

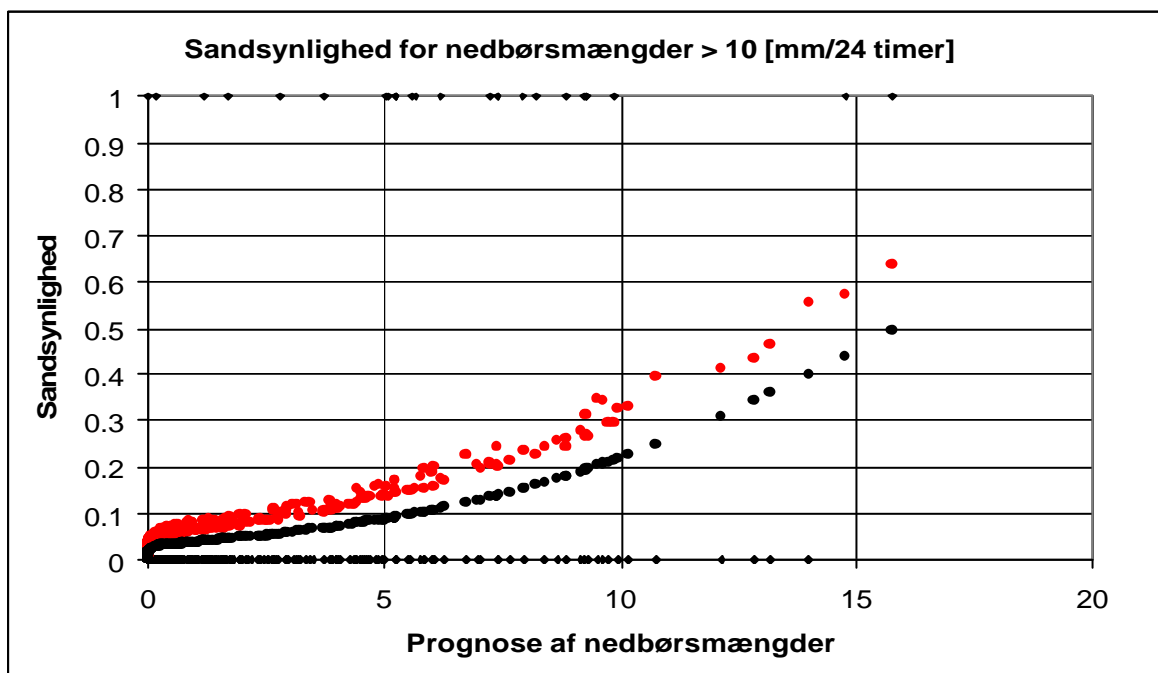
Ligesom ved traditionel lineær regression af en meteorologisk parameter kan koefficienterne, a_i , i regressionsligningen bestemmes rekursivt ved hjælp af et logistisk Kalman filter, når de observeret Y 'er bliver tilgængelige. De eneste ændringer i forhold til et traditionelt Kalman filter er, at estimatet \hat{Y} transformeres med den logistiske funktion $P(\hat{Y})$ og at fejlen $(Y - \hat{Y})$ ændres til $2 * \text{Arctanh}(I-P)$. $\text{Arctanh}(I-P)$ varierer mellem $+$ og $-$ uendelig, når $I-P$ varierer mellem -1 og 1 .

Det har desuden vist sig nødvendigt at indføre et korrektionsmaksimum givet ved, at fejlsandsynligheden I-P skal være mindre end en tærskelværdi. Anvendelse af et korrektionsmaksimum har primært vist sig nødvendigt ved store nedbørsmængder.

Et Kalman filter er afhængig af observations og system variansen, som begge skal bestemmes på forhånd.

Observations variansen er blevet bestemt som middelværdien $(I-P_{\text{klima}})^2$ bestemt ud fra en delmængde af data materialet og hvor P_{klima} er sandsynligheden for nedbørsmængder større end den aktuelle tærskelværdi bestemt ud fra samme delmængde af data materialet. Som vil fremgå senere laves verifikation på den resterende del af data sættet.

For regression med n uafhængige variable er system variansen en $n \times n$ matrice. Bestemmelse af system variansen er gjort ved at initialisere system variansen til 0 i alle matrice elementerne og derefter beregne opdatering for 20 dage. Kalman filteret bestemmer da selv en matrice til opdatering af koefficienterne. Det antages nu at denne matrice kan bruges til at vurdere det indbyrdes størrelsesforhold af koefficienterne i system varians matricen. De aktuelle værdier fås ved at skalere så den geometriske middel af diagonal elementerne er 1 (dvs. at n'te rod af produktet af diagonalelementer er 1) og derefter skalere med R/N^2 , hvor R er observations variansen og N er systemhukommelsen målt i dage. Der er udført forsøg med forskellige antal dage for systemhukommelse. N = 30 dage giver fornuftige resultater, men kan dog varieres for de forskellige filtre.



Figur 6 Sandsynligheden for nedbørsmængder større end 10 [mm/24 timer] som funktion af den kategoriske prognose for nedbørsmængder. Røde prikker er Kalman filter og sorte prikker er logistisk regression.

I figur 6, som svarer til figur 5 for logistisk regression, viser de røde prikker sandsynligheden for nedbørsmængder større end 10 [mm/24 timer] beregnet med Kalman filteret og de sorte den tilsvarende sandsynlighed beregnet med logistisk regression. Sandsynligheder

beregnet med Kalman filteret ligger lidt højere end sandsynligheder beregnet med logistisk regression. Det skyldes at koefficienterne i logistisk regression er beregnet på en delmængde af data sættet og derefter anvendt på resten af data sættet, hvorimod et Kalman filter netop bestemmes løbende ud fra de nyeste data.

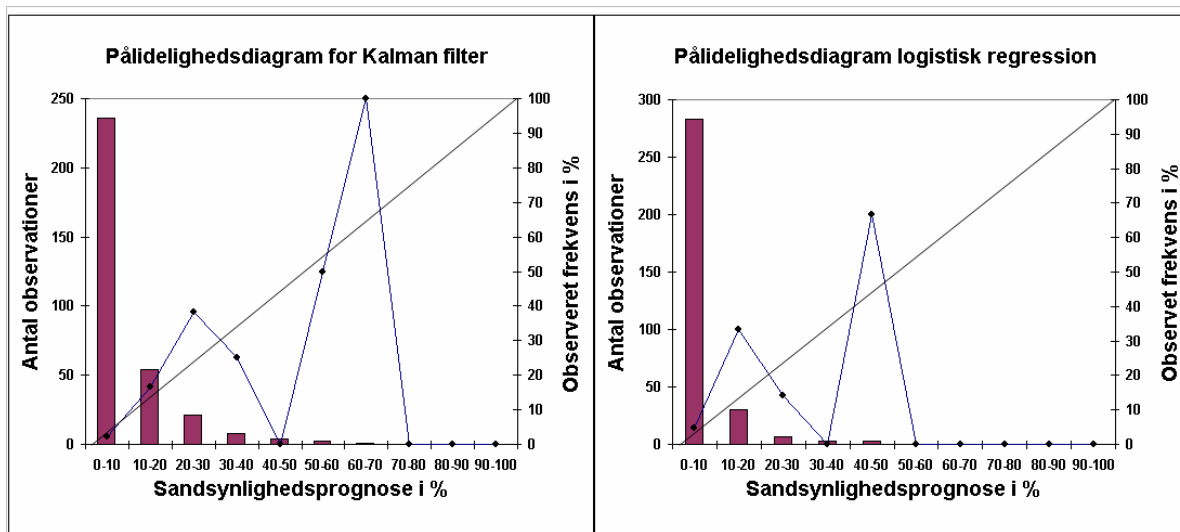
Hvorledes disse to metoder kan verificeres vil blive gennemgået i næste kapitel.

6. Verifikation af sandsynlighedsprognoser.

6.1 Pålidelighed og skarphed

Et krav til sandsynlighedsprognoser, er at de skal være pålidelige. Pålidelighed udtrykker, at en hændelse forudsagt med sandsynligheden P også i praksis skal realiseres med frekvensen P . Sandsynlighedsprognoserne grupperes derfor i 10 grupper svarende til sandsynlighedsintervallerne 0-10%, 10-20%, 20-30%, 30-40%, 40-50%, 50-60%, 60-70%, 70-80%, 80-90%, 90-100%. For hvert interval er antallet af tilfælde, hvor sandsynlighedsprognosen er faldet indenfor intervallet bestemt. Ligeledes er for hvert interval antallet af gange, hvor sandsynlighedsprognosen er faldet indenfor intervallet og der tillige samtidigt blev observeret en nedbørsmængde større end tærskelværdien bestemt. Forholdet mellem disse to tal er da frekvensen af observerede hændelser i intervallet.

Figur 7 viser et pålidelighedsdiagram for et Kalman filter og for logistisk regression for nedbørsmængder større end 10 [mm/24 timer]. Et fuldstændigt pålideligt system vil resultere i en kurve, hvor sandsynligheden er lig med den observerede frekvens, hvilket svarer til en ret linie gennem (0,0) og (100,100). De store udsving skyldes til dels det lille antal hændelser i data materialet med nedbørsmængder større end 10 [mm/24 timer].

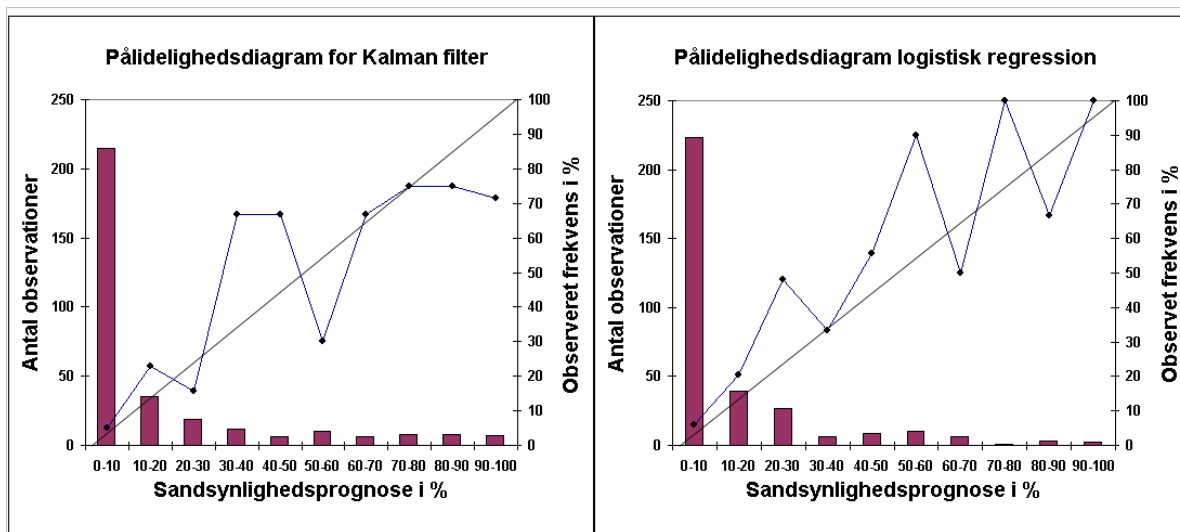


Figur 7 Pålidelighedsdiagrammer for Kalman filter og logistisk regression for nedbørsmængder større end 10 [mm/24 timer]

Figur 8 viser tilsvarende pålidelighedsdiagrammer for nedbørsmængder større end 5 [mm/24 timer]. Her ses betydeligt mindre udsving i kurverne.

Foruden pålidelighedskurverne i figur 7 og 8, er ligeledes for hvert interval antallet af gange hvor sandsynlighedsprognosen er faldet indenfor intervallet vist som søjler med skalaen til venstre i diagrammet. For nedbørsmængder større end 10 [mm/24 timer] er antallet af sandsynlighedsprognoser med sandsynlighed mellem 0 og 10% langt over 200, mens der ikke er ret mange sandsynlighedsprognoser med store sandsynligheder. Det skyldes dels igen det

lille antal hændelser med nedbørsmængder større end 10 [mm/24 timer], men også prognosen relativt ringe evne til at forudsige nedbørsmængder større end 10 [mm/24 timer].



Figur 8 Pålidelighedsdiagrammer for Kalman filter og logistisk regression for nedbørsmængder større end 5 [mm/24 timer]

Da pålideligheden for sandsynlighed mellem 0 og 10% er stor, er sandsynlighedsprognoserne er altså relativt gode til at forudsige, når der ikke kommer nedbørsmængder større end 10 [mm/24 timer].

For nedbørsmængder større end 5[mm/24 timer] er antallet af sandsynlighedsprognoser med sandsynlighed mellem 0 og 10% stadig mere end 200, men nu er der flere hændelser med sandsynlighedsprognoser med større sandsynligheder.

Udover pålidelighedsdiagrammer anvendes også Brier skill score, der i en vis forstand tager hensyn til både pålidelighed og skarphed og giver mulighed for en direkte sammenligning af to metoder.

6.2 Brier score og skill score

Brier score er summen af kvadratet på fejlsandsynligheden (I-P), hvor I er indikatorfunktion, som er lig med 1, hvis der er observeret en nedbørsmængde større end tærskelværdien og nul ellers. Brier score er normeret til 1 for en fuldstændig fejlagtig prognose, hvor sandsynlighedsprognosen er lig med 0, hvis der den pågældende dag er observeret en nedbørsmængde større end tærskelværdien, og ellers 1. Den perfekte prognose har en Brier score på 0.

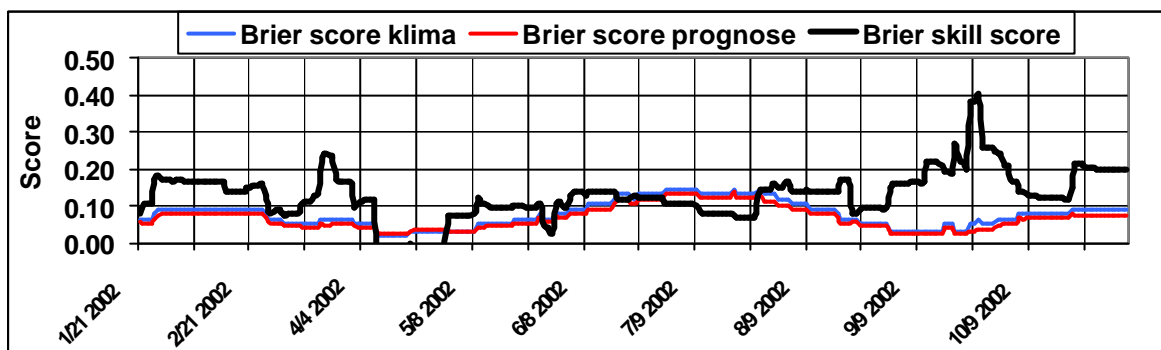
Skill score er generelt defineret som forskellen i score mellem prognosen og en reference prognose, relativt til forskellen i score mellem den perfekte prognose og reference prognosen. Skill er dermed et udtryk for hvor mange procent af den maksimalt mulige forbedring af reference prognosens skill, den aktuelle prognose giver.

Skill scores er 1 for perfekte prognoser, nul hvis prognosen er neutral i forhold til reference prognosen og negativ hvis prognosen er ringere end reference prognosen. Som reference prognose vælges den klimatiske sandsynlighed P_{klima} , som bestemmes ud fra en delmængde af data sættet. Scores kan derefter beregnes for resten af data sættet.

For Brier score bliver Brier skill score $BSS = (BS_{\text{ref}} - BS_{\text{prognose}}) / BS_{\text{ref}}$, da Brier score for den perfekte prognose er nul. BS_{ref} er Brier score for reference prognosen og BS_{prognose} er Brier score for prognosen.

Brier skill score for den logistiske regression er 0,11 og for kalman filteret 0.13, altså lidt bedre end den logistiske regression. Det skyldes nok som tidligere nævnt, at koefficienterne i logistisk regression er beregnet på en delmængde af data sættet og derefter anvendt på resten af data sættet, hvorimod et Kalman filter netop bestemmes løbende ud fra de nyeste data.

Brier score og Brier skill score beregnes ofte løbende. Figur 9 viser 60 dages løbende Brier score og Brier skill score for tærskelværdien 10 [mm/24 timer].



Figur 9. 60 dages løbende Brier score og Brier skill score for nedbørsmængder > 10 [mm/24 timer]

Her ses, at der er størst skill i vinter og efterårs månederne, hvor der er mest nedbør. Der er imidlertid relativt lidt skill i sandsynlighedsprognoserne. I næste kapitel vises, hvorledes sandsynlighedsprognoserne kan anvendes så de alligevel giver en forbedring i forhold til for eksempel kategoriske prognoser.

6.3 Cost/Loss Value

Udover pålidelighedskurver og Brier score, som begge er objektive mål, anvendes også Cost/Loss Value og Relative Operational Characteristic, ROC som verifikations mål. Cost/Loss Value er et subjektivt mål baseret på, at der skal træffes en beslutning af en konkret bruger, og derigennem afhængig af den enkelte brugers behov. Det medfører til gengæld, at en prognose, som for eksempel har en relativt lille Brier skill score, alligevel for nogen brugere har en vis værdi.

En brugerbeslutning kan bestå i at afgøre, om en aktivitet skal beskyttes mod for eksempel store nedbørsmængder, fordi der vil være et tab (loss), hvis der kommer store nedbørsmængder uden at aktiviteten er beskyttet. Beskyttelse af aktiviteten har imidlertid også omkostninger. Der skal altså på baggrund af den tilgængelige viden træffes beslutninger så

risikoen for unødige omkostninger bliver mindst mulige. Den tilgængelige viden er ikke er fuldstændig, idet kun prognoser med en hvis usikkerhed kendes.

I det følgende antages som eksempel at 10 [mm/24 timer] er en kritisk nedbørsmængde for en tænkt brugers aktivitet. I dette simple eksempel er det altså en binær beslutning om beskyttelse/ej beskyttelse, som prognoserne skal bidrage med information til.

Hvis kun den klimatologiske hyppighed er kendt, hvilket svarer til en klimatologiske sandsynlighedsprognose, kan der kun træffes en overordnet beslutning, nemlig enten altid at beskytte eller aldrig at beskytte, fordi der ikke er nogen variation i prognosen.

Beslutningen skal træffes, så der opnås minimale omkostninger. Antag at tabet er L kr (Loss), hvis aktiviteten udføres og der kommer mere end 10 [mm/24 timer] og antag, at det koster C kr (Cost) at beskytte aktiviteten mod nedbør.

Beslutningen viser sig at være afhængig af forholdet C/L. Omkostningen over N dage er nemlig, hvis der altid beskyttes $N \cdot C$ og, hvis der aldrig beskyttes $N \cdot (\text{den klimatologiske hyppighed}) \cdot L$. Det er altså billigst altid at beskytte hvis $C/L < (\text{den klimatologiske hyppighed})$.

Den klimatologiske prognose er det mindste grundlag for beslutning og bruges derfor som reference prognose.

Foruden den klimatologiske prognose findes de kategoriske prognoser og sandsynlighedsprognoserne, samt en tænkt perfekt prognose. Cost/Loss Value er en vurdering af den værditilvækst(besparelse) disse prognoser kan give i forhold til den klimatologiske prognose.

6.3.1 Kategoriske prognoser

Den kategoriske prognose giver en værdi for den akkumulerede nedbør over 24 timer. Beslutningen bliver derfor at beskytte, hvis prognosen er over 10 [mm/24 timer] og ellers ikke. Omkostningerne beregnes ud fra en tilhørende kontingenstabel vist i tabel 4, hvor a, b, c og d er normerede frekvenser så $a + b + c + d = 1$.

Tabel 4 Kontingens tabel med frekvenser. $a+b+c+d=1$, klimatologisk hyppighed = $b + d$.

Der beskyttes, når prognosen er Ja og ellers ikke. Omkostningerne over N dage til beskyttelse er: $N \cdot (c + d) \cdot C$. Hvis der ikke beskyttes og der alligevel kommer en nedbørsmængde større end 10 [mm/24 timer] bliver omkostningen (tabet): $N \cdot b \cdot L$. Den samlede omkostning, O er da summen af disse to:

$$O = N * (c + d) * C + N * b * L \quad \text{ligning (1)}$$

Hvis den klimatologiske hyppighed = $(b + d)$ betegnes KL , kan frekvensen af detekteret hændelse defineret ved $FOD = d/(b + d)$ bestemmes som d/KL . Frekvensen af falske alarmer givet ved $FAR = c/(a + c)$ kan udtrykkes som $c/(1 - KL)$, hvor en falsk alarm er de tilfælde, hvor prognosen lover mere end 10 [mm/24 timer] uden at dette indtræffer.

Det bør dog lige bemærkes her, at frekvensen af falske alarmer her er relativt til observationerne, hvor man ofte måler frekvensen af falske alarmer relativt til prognosen som $c/(c + d)$.

Af ovenstående fås at:

$$b = KL * (1 - FOD)$$

$$c = FAR * (1 - KL)$$

$$d = FOD * KL$$

Den samlede omkostning for den kategoriske prognose kan da udtrykkes ved KL , FAR og FOD ved at erstatte b , c og d i ligning (1):

$$O = \{(1 - KL) * FAR * C - KL * FOD * (L - C) + KL * L\} * N \quad \text{ligning (2)}$$

Hvis omkostningen ved anvendelse af den kategoriske prognose er mindre end den tilsvarende for den klimatologiske prognose, som er minimum af $(N * C)$ og $(N * KL * L)$, har man opnået en værditilvækst ved at anvende den kategoriske prognose. Værditilvæksten afhænger altså ikke og alene af prognosens FOD og FAR , men også af brugerens C og L værdier.

Den perfekte prognose repræsenterer den minimalt opnåelige omkostning. Den perfekte prognose har $FOD=1$ og $FAR=0$. Omkostningen for den perfekte prognose findes ved indsættelse af $FOD=1$ og $FAR=0$ i ligning (2) og bliver dermed $(N * KL * C)$.

I lighed med definitionen af skill score defineres Cost/Loss Value som forholdet mellem værditilvæksten for prognosen i forhold til reference prognosen og værditilvæksten for den perfekte prognose i forhold til reference prognosen. Cost/Loss Value er altså forholdet mellem den opnåede værditilvækst ved anvendelse af prognosen i forhold til den maksimale opnåelige værditilvækst. Cost/Loss Value beregnes som:

$$\text{Cost/Loss Value} = \frac{(1-KL)*FAR*C/L - KL*FOD*(1-C/L) + KL - \min(C/L, KL)}{KL*C/L - \min(C/L, KL)} \quad \text{ligning (3)}$$

For et givet prognosesystem kan Cost/Loss Value for brugere med forskellige C/L tegnes i et diagram. Maksimum fås når C/L er lig med den klimatologiske hyppighed, og værditilvæksten her er $KS = FOD - FAR$, også kaldet Kuipers skill score. Cost/Loss Value vises i figur 10 sammen den tilsvarende kurve for sandsynlighedsprognoser.

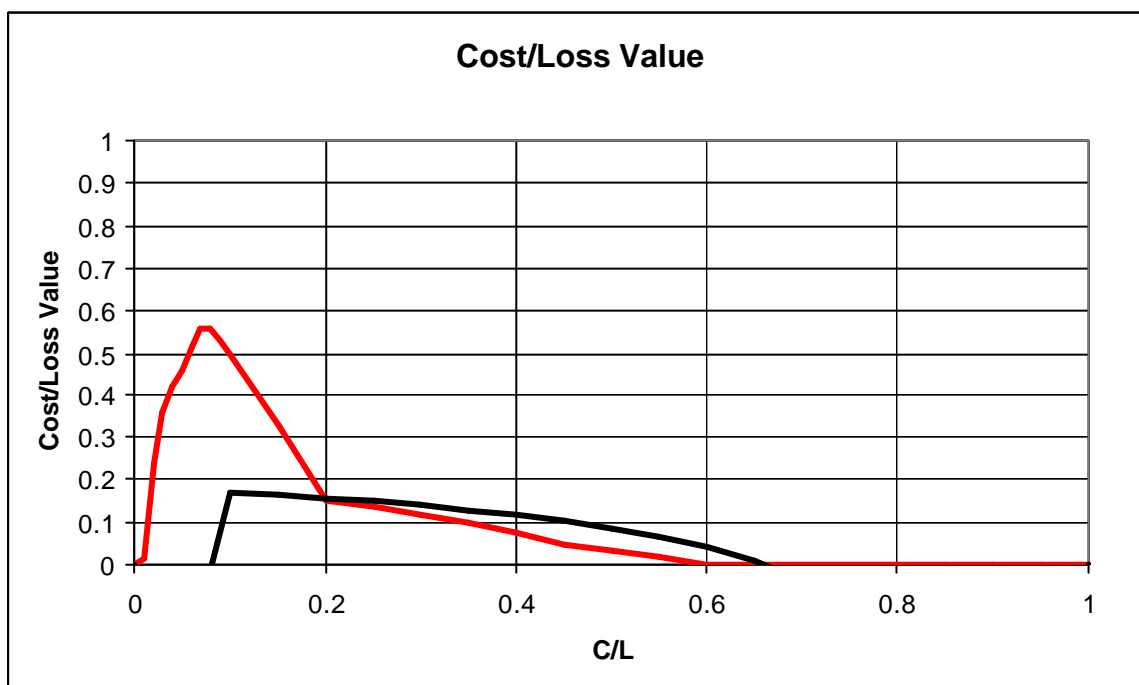
6.3.2 Sandsynlighedsprognoser

Hvis sandsynlighedsprognoser er til rådighed, bliver spørgsmålet, over hvilket mindste sandsynlighedsniveau det er optimalt at beslutte og beskytte, og hvad værditilvæksten så bliver. Altså hvis sandsynligheden er over en hvis tærskelværdi beskyttes ellers ikke. For en

given tærskelværdi for sandsynligheden kan en kontingenstabel bestemmes på samme måde som for den kategoriske prognose og de tilhørende FAR og FOD bestemmes. Cost/Loss Value kan så beregnes for alle C/L.

Ud fra et sæt af stigende tærskelsandsynligheder bestemmes ved hjælp af de tilhørende kontingenstabeller et sæt af FAR og FOD. Disse definerer så et helt sæt af kurver for Cost/Loss Value som funktion af C/L. Indhyldningskurven for dette sæt af kurver er da sandsynlighedsprognosernes totale Cost/Loss Value for brugere med forskellige C/L forhold.

Figur 10 viser Cost/Loss Value kurver for den kategoriske prognose (sort) og for sandsynlighedsprognose (rød). For et lille C/L forhold har sandsynlighedsprognose langt den største værditilvækst i forhold til den tilhørende kategoriske prognose. Det er en stor fordel, da de fleste brugere vil have et lille C/L forhold, altså små omkostninger til beskyttelse og store tab ved at undlade og beskytte når hændelsen indtræffer. For stort C/L forhold



Figur 10. Cost/Loss Value kurver for kategoriske prognoser (sort) sandsynlighedsprognoserne (rød).

har ingen af prognoserne nogen værditilvækst, fordi det her stort set er ligegyldigt om der bliver beskyttet eller ej, da omkostningen ved beskyttelse er det samme som tabet ved ikke at beskytte.

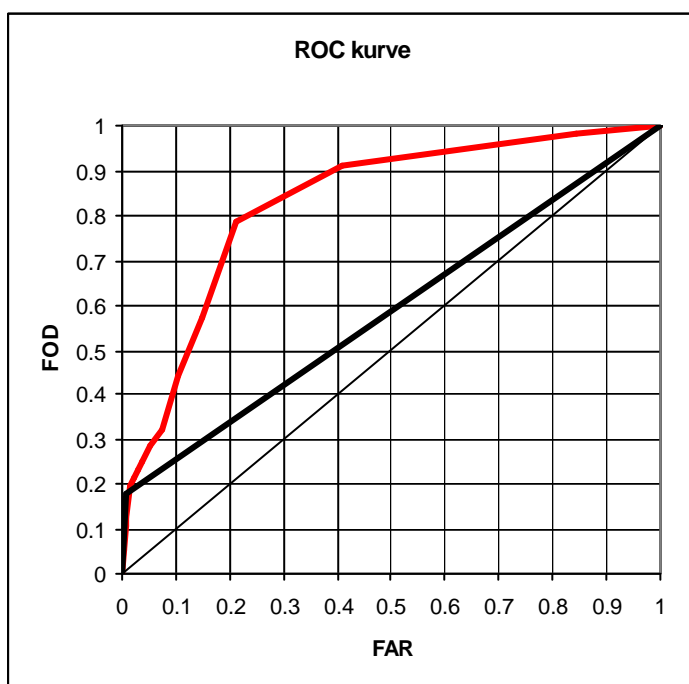
Ved store C/L forhold giver i dette eksempel den kategoriske prognose den største værditilvækst.

6.4 ROC kurver

FAR og FOD for forskellige tærskelværdier for sandsynlighed kan afbildes i et diagram som en såkaldt ROC-kurve gennem (0,0) og (FAR,FOD) og (1,1). Den perfekte prognose har kun et sæt (FAR,FOD) nemlig, FAR=0 og FOD=1, ingen falske forudsigelser og alle hændelser forudsiges korrekt. Den klimatologiske prognose har FAR=FOD=0 svarende til aldrig at forudsige hændelsen og dermed heller ingen falske eller FAR=FOD=1 svarende til altid at forudsige hændelsen, hvorved alle ikke hændelser bliver forudsagt falskt.

Den klimatologiske prognoses ROC-kurve er den rette linie gennem (0,0) og (1,1). Arealet under denne kurve er $\frac{1}{2}$. Arealet under ROC-kurven for den perfekte prognose, som går gennem (0,0), (0,1) og (1,1) er 1. Arealet under en given ROC-kurve kan derfor anvendes som skill for et givet prognosesystem. Hvis der er skill i systemet, skal arealet være så stort som muligt og større end $\frac{1}{2}$, som svarer til ingen mulighed for værditilvækst for nogen brugere.

Figur 11 viser ROC-kurverne for den kategoriske prognose (sort) og sandsynlighedsprognoserne (rød). Her ses tydeligt at sandsynlighedsprognoser har et klart større areal end den kategoriske prognose.



Figur 11 ROC-kurverne for den kategoriske prognose (sort) og sandsynlighedsprognoserne (rød).

7. Konklusion

DMI-HIRLAM's kategoriske prognoser er blevet analyseret. Der er opstillet en metode til bestemmelse af sandsynlighedsprognoser ved hjælp af et logistisk Kalman filter. Forskellige verifikationsmetoder af sandsynlighedsprognoser er blevet beskrevet i detaljer.

Analysen af DMI-HIRLAM's kategoriske nedbørsprognoser viser, at disse underestimerer store nedbørsmængder. Sandsynlighedsprognoser baseret på logistisk Kalman filtrering af DMI-HIRLAM's kategoriske nedbørsprognoser forbedrer mulighederne for anvendelse, idet Cost/Loss Value for brugere med lille Cost/Loss forhold er væsentligt større end det tilsvarende Cost/Loss Value for den kategoriske prognose, hvilket yderligere illustreres ved et større areal under ROC-kurven.

De nuværende sandsynlighedsprognoser er gode til at forudsige, hvornår der ikke kommer store nedbørsmængder. Altså mange pålidelige tilfælde med lille sandsynlighed for store nedbørsmængder.

Forbedringer af DMI-HIRLAM vil også forbedre sandsynlighedsprognoserne for kraftig nedbør. Det forventes at disse forbedringer også vil øge muligheden for at forudsige større sandsynligheder for store nedbørsmængder.