BILAGA II

Extremvärdesstatistik och osäkerhet

I denna något ”spretiga” bilaga har samlats ett antal sektioner som beskriver och fördjupar olika metoder och resultat kopplade till den statistiska bearbetningen av extremvärden, inklusive osäkerhet.

Bilaga II.1 Om återkomsttider

Ett vedertaget sätt att modellera extrem nederbörd är med *återkomsttider* (Coles, 2001). Återkomsttider är ett mått på hur extremt ett värde är, och kan intuitivt tolkas som med hur många års mellanrum som händelsen i *genomsnitt* inträffar. Här är ordet *genomsnitt* väldigt viktigt, en 100-årshändelse kan inträffa två år i rad, men i långa loppet är det alltså i genomsnitt 100 år mellan dessa händelser.

Den vanligaste ansatsen vid arbete med återkomsttider är att bygga en modell kring nederbördsseriens årshögsta värden. Ansatsen har även tillämpats i denna studie. Värdena antas vara oberoende och följa en och samma sannolikhetsfördelning. Utifrån denna fördelning får man kunskap om hur årets högsta nederbörd beter sig på den aktuella platsen.

I projektet har två olika metoder använts för att beräkna återkomsttider. Metoderna är Årsmax-metoden och Peak over Threshold (POT). De är beskrivna i respektive delavsnitt nedan

Statistikteorin som återkomsttider bygger på kallas extremvärdesteori. Den viktigaste satsen inom denna är *extremvärdessatsen* som, under vissa förutsättningar, tillåter antagandet att årsmax-värdena (årsmax = högsta värdet under året) följer en viss sannolikhetsfördelning. I princip måste årsmax-värdena vara oberoende och likafördelade (dvs. årsmax år 1900 bör ”bete sig” som årsmax år 1990).

Det är inte nödvändigt att använda kalenderår då de mest extrema händelserna extraheras. Generellt behöver tidsperioden indelas i block och sedan hämtas det högsta värdet inom varje block. Blocken konstrueras så att det högsta värdet inom varje block är oberoende av de andra blockens högsta värden, och att de alla följer samma fördelning.

Det är viktigt att tolka återkomsttider korrekt. Exempelvis ska100-årsnederbörden tolkas som den nederbörd som har 1 chans på 100, varje år, att det överskrids. På samma sätt ska man tolka meningen ”Nederbörden 40 mm har återkomsttid 50 år” som att det varje år är 1 chans på 50 att nederbörden överskrider 40 mm.

Sannolikheten under en längre period är därmed inte densamma som för ett enskilt år. Den ackumulerade sannolikheten för att en händelse sker under en längre tidsperiod visas för olika återkomsttider i tabell 1. För vidare fördjupning se Blom m.fl. (2005) eller Coles (2001).

Tabell 1. Sannolikheten att en händelse med en viss återkomsttid överskrids minst en gång under en given period.

Återkomsttid (år)	Sannolikhet under 100 år (%)	Sannolikhet under 200 år (%)	Sannolikhet under 300 år (%)
50	87	98	100
100	63	87	95
200	39	63	78
300	28	49	63
1000	10	18	26
10 000	1	2	3

Bilaga II.2 Årsmaxmetoden (AM)-metoden

I årsmaxmetoden delas tidsserien i ”block”, vanligen om ett kalenderår per block, och det högsta värdet extraheras från varje block. Då erhålls en dataserie med årsmax-värden. Dessa värden anpassas sedan till en sannolikhetsfördelning.

Metoden kallas även ofta för *årsmax*-metoden, eftersom block om ett år används.

Dataserien behöver inte nödvändigtvis indelas i block över kalenderår. Blocken ska konstrueras på sådant sätt att det högsta värdet inom varje är oberoende av de andra blockens högsta värden, och att de alla följer samma fördelning (Coles, 2001).

Låt den sannolikhetsfördelning som anpassats till data ha fördelningsfunktion $F(x;\theta)$, där x är datavektorn (t.ex. årsmax-värden), och θ är fördelningens parametervektor. Återkomstnivån R för återkomsttid T år kan beräknas genom

$$R = F^{-1}(1-\frac{1}{T};\theta)$$

Med andra ord är R det värde där fördelningsfunktionen F antar värdet $1-\frac{1}{T}$.

Vi kan kasta om lite i ekvationen ovan för att få ett uttryck för återkomsttiden givet nivån R:

$$T = \frac{1}{1-F(R;\theta)}$$

Extremvärdessatsen ger stöd till användandet av GEV-fördelningen, samt dess specialfall Gumbel (vilket man får när GEV:s formparameter = 1). I praktiken kan dock godtycklig sannolikhetsfördelning användas för modellering av extremvärden, så länge som den är kontinuerlig och har stöd för de värden som extremvärdena kan anta. I detta projekt har följande sannolikhetsfördelningar använts för att beräkna återkomsttider med årsmaxmetoden:

Generalized Extreme Value (GEV)

$F(x;\mu,\sigma,\xi) = \exp(-(1+\xi*(\frac{x-\mu}{\sigma})^{-1/\xi})$

Stödet är $x \in \mathbb{R}$

μ är platsparametern (”location”)

σ är skalparametern (”scale”)

ξ är formparametern (”shape”)

Gumbel

$z = \frac{(x-\mu)}{\sigma}$

$F(x;\mu,\sigma) = 1 - \exp(-\exp(z))$

Stödet är $x \in \mathbb{R}$

μ är platsparametern (”location”)

σ är skalparametern (”scale”)

Log-Pearson typ III

Låt x vara tidsserien med totalt n årsmaxvärden. x_i är värdet på plats $i, i = 1, \dots, n$.

Skapa

$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$

$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2}$

$s_1 = \sum_{i=1}^{n} x_i$

$s_2 = \sum_{i=1}^{n} {x_i}^2$

$s_3 = \sum_{i=1}^{n} {x_i}^3$

$t_1 = n^2 * s_3$

$t_2 = -3 * n * s_1 * s_2$

$t_3 = 2 * {s_1}^3$

$u = n * (n-1) * n-2) * \sigma^3$

$C = \frac{t_1+t_2+t_3}{u}$

Låt p vara CDF-värdet av intresse (För T-årsvärdet så sätts p=1-1/T), och $F^{-1}(p)$ är återkomstnivån för detta CDF-värde. Formeln för $F^{-1}(p)$ ser ut som följer:

$F^{-1}(p) = \exp(\mu + \mathrm{K(p, C)} * \sigma)$

Där $\mathrm{K(p, C)}$ beräknas enligt nedanstående steg

$w = \sqrt{\log(\frac{1}{p^2})}$

$a_0 = 2.515517$

$a_1 = 0.802853$

$a_2 = 0.010328$

$b_0 = 1$

$b_1 = 1.432788$

$b_2 = 0.189269$

$b_3 = 0.001308$

$z = w - \frac{a_0 + a_1 * w + a_2 * w^2}{b_0 + b_1 * w + b_2 * w^2 + b_3 * w^3}$

$k = C/6$

$t_1 = z$

$t_2 = (z^2 - 1) * k$

$t_3 = \frac{(z^3 - 6 * z) * k^2}{3}$

$t_4 = (z^2 - 1) * k^3$

$t_5 = z * k^4$

$t_6 = \frac{k^5}{3}$

$K = t_1 + t_2 + t_3 + t_4 + t_5 + t_6$

Log-Pearson typ III har använts i t.ex. Bilaga III för beräkning av återkomsttider för skyfall.

Bilaga II.3 Peak-Over-Threshold (POT)-metoden

Peak over threshold (POT) är en metod för att beräkna återkomsttider från en tidsserie (Coles, 2001). Grundprincipen är att oberoende händelser extraheras över en viss tröskel, och de anpassas sedan till en sannolikhetsfördelning som återkomsttider kan beräknas utifrån.

För de utvalda tidsserierna har en POT-analys utförts. Först extraherades händelser över en viss fix tröskel. Enligt Pickands-Balkema-de Haans sats (Pickands, 1975) gäller att om händelserna över tröskeln är oberoende och likafördelade, vilket de bör vara om tröskeln valts på ett klokt sätt, så kommer de att följa en Generaliserad Pareto (GP) fördelning. Denna fördelning kan sägas vara skräddarsydd för POT-metoden. Parametrarna av denna fördelning har skattats med ML-metoden (se avsnitt 3.3.1).

Hur tröskeln i POT-analysen väljs är ett aktivt forskningsområde och det finns ofta inget uppenbart svar. Om tröskeln väljs för hög så är man mer garanterad att händelserna är oberoende och verkligen relevanta för extremvärdesanalysen, men man kan då gå miste om relevanta händelser. Om tröskeln väljs för låg så får man in mycket brus i data, dvs. händelser som inte är extrema och därmed inte av intresse för analysen.

För stationsårsmetoden har tröskeln valts så att man får lika många händelser som antal år i datamaterialet, d.v.s. om tidsserien är X år lång väljer man den tröskel som ger X händelser som överskrider den. Anledningen till detta var för att få den jämförbar med årsmaxmetoden.

Generalized Pareto (GP)

Fördelningsfunktionen för GP-fördelningen är

$$F(\xi,\sigma)(x) = \begin{cases} 1-\left(1+\frac{\xi x}{\sigma}\right)^{-\frac{1}{\xi}} & \text{för } \xi \neq 0 \\ 1-e^{-\frac{x}{\sigma}} & \text{för } \xi = 0 \end{cases}$$

Stödet är $x > 0$

Återkomsttider med POT-metodik beräknas på följande sätt:

Låt F vara fördelningsfunktionen av den skattade GP-fördelningen. Denna funktion är inverterbar (då den är strängt växande), och vi kan kalla dess invers F^{-1}. Antag att T-årsnivån söks, där T är återkomsttiden i år (till exempel 100-årsnivån).

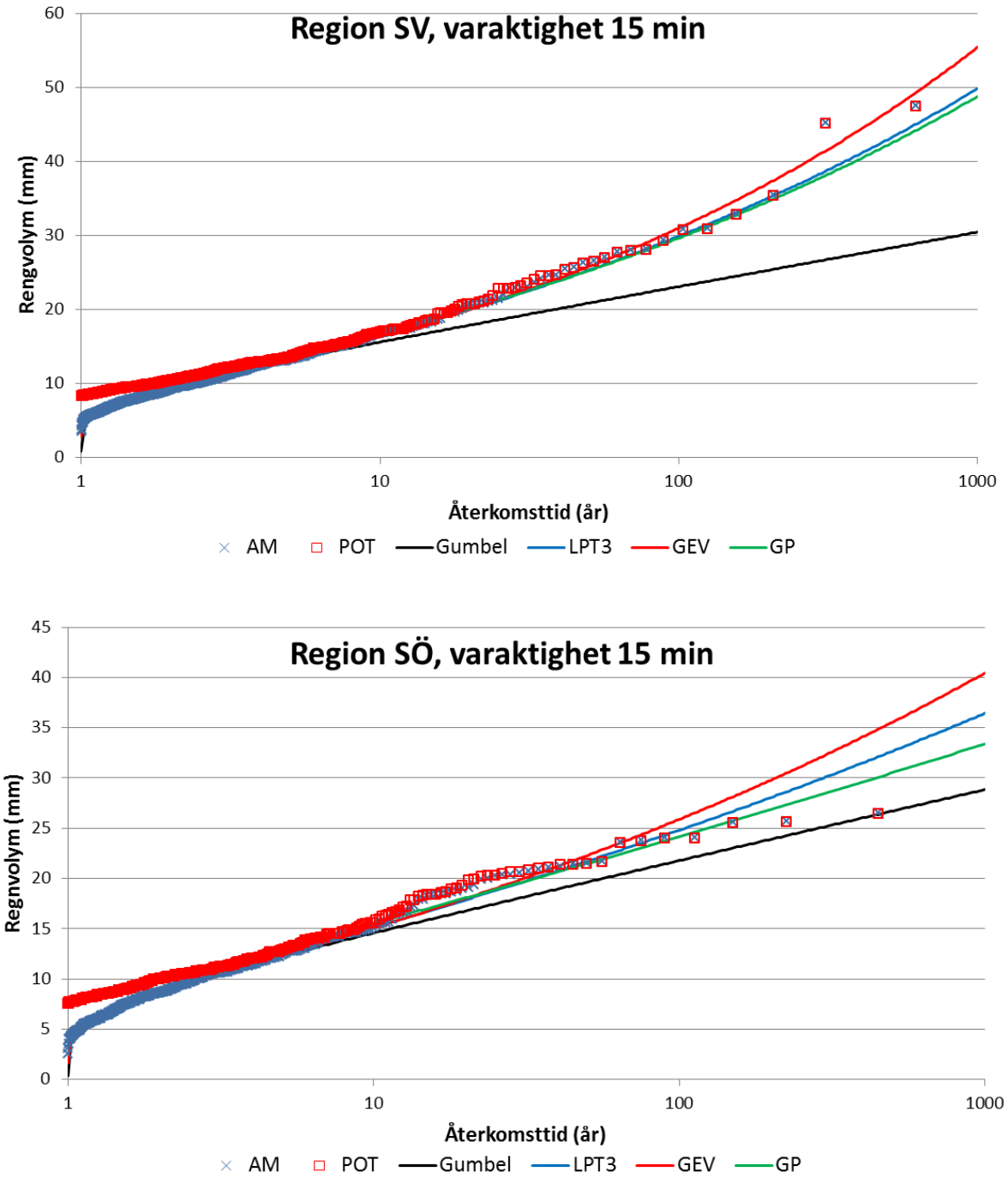
För att kunna uttrycka återkomsttider i denna enhet behöver hänsyn tas till antalet händelser per år. Därför skalas sannolikheten för överskridande, $1-\frac{1}{T}$, med genomsnittligt antal händelser per år enligt

$$P = (1-\frac{1}{T})^{\frac{1}{genomsnittligt\ antal\ händelser\ per\ år}}$$

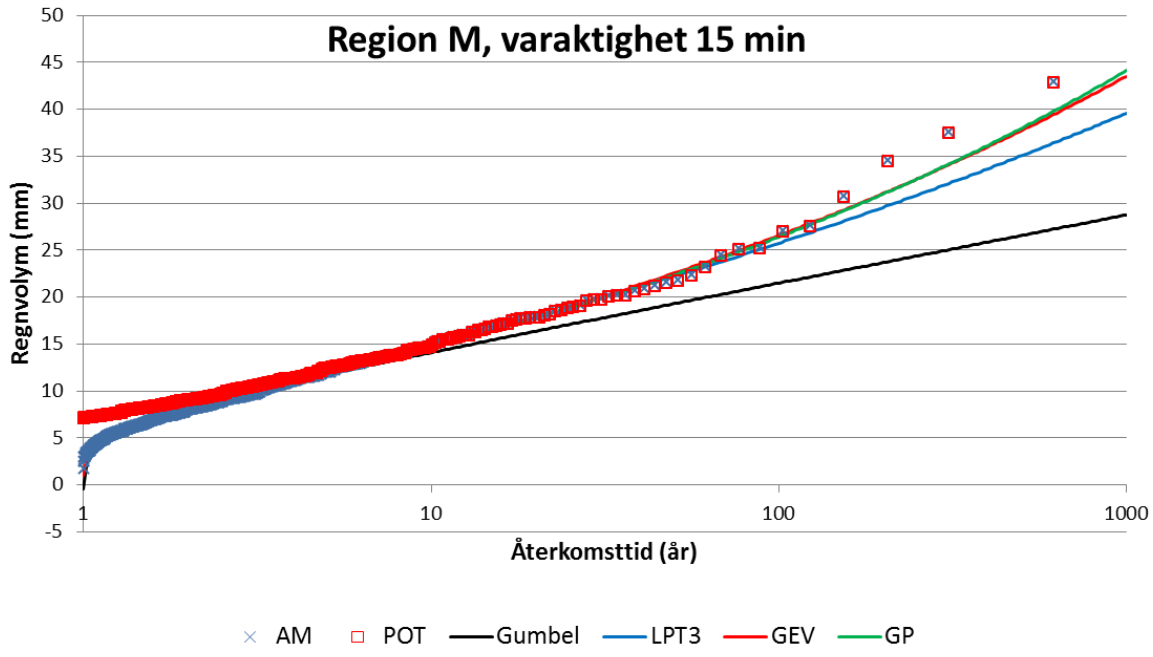
Genomsnittligt antal händelser per år är $\frac{Totalt\ antal\ händelser\ över\ tröskeln}{Antal\ år\ med\ data}$.

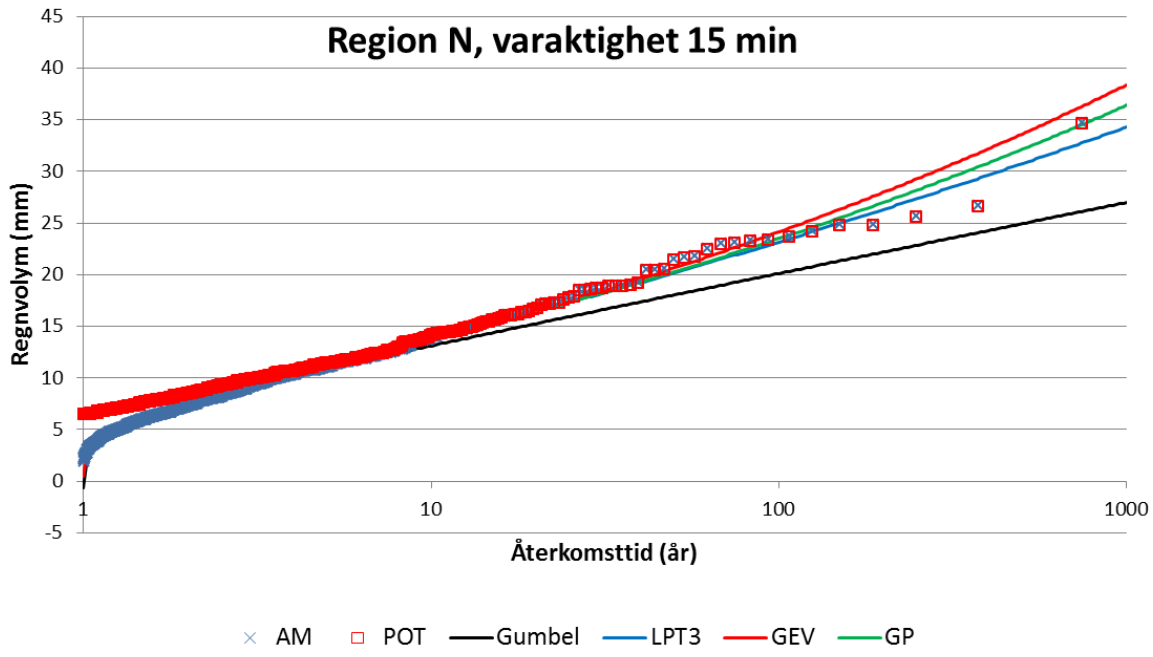
Återkomstnivån Rför återkomsttid T är sedan $R = F^{-1}(P)$.

Bilaga II.4 Anpassningar för olika regioner och varaktigheter

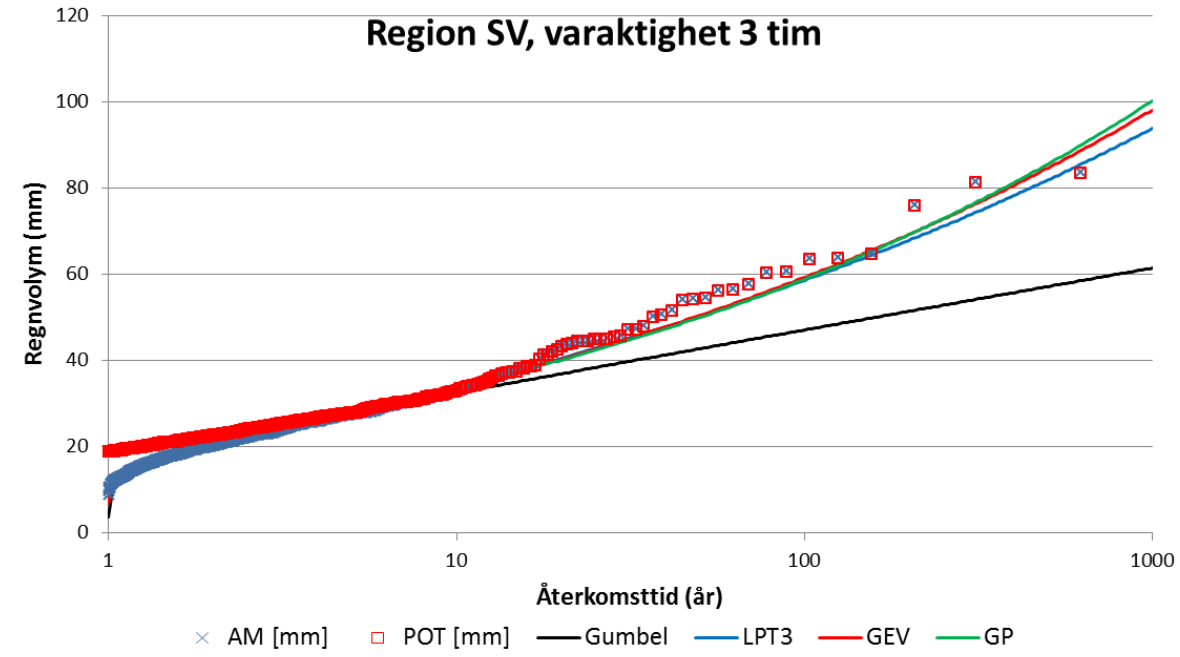


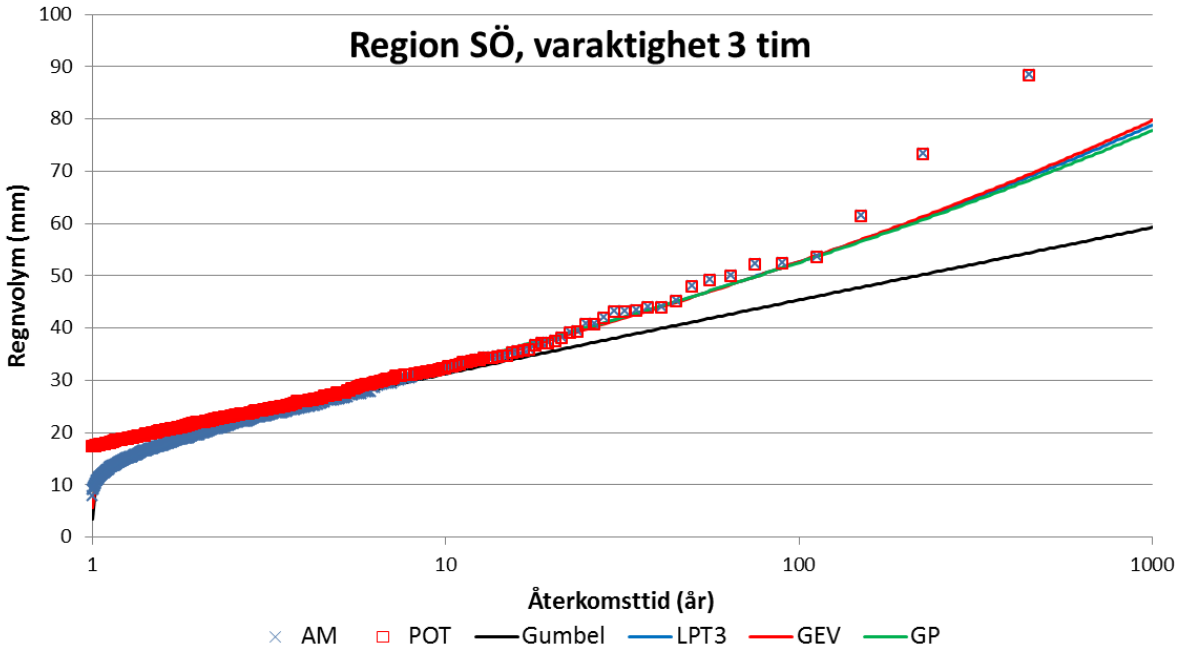
Figur 1. Framtagna 15-min extremvärden och anpassade sannolikhetsfördelningar för de sammanslagna serierna i region SV och SÖ.



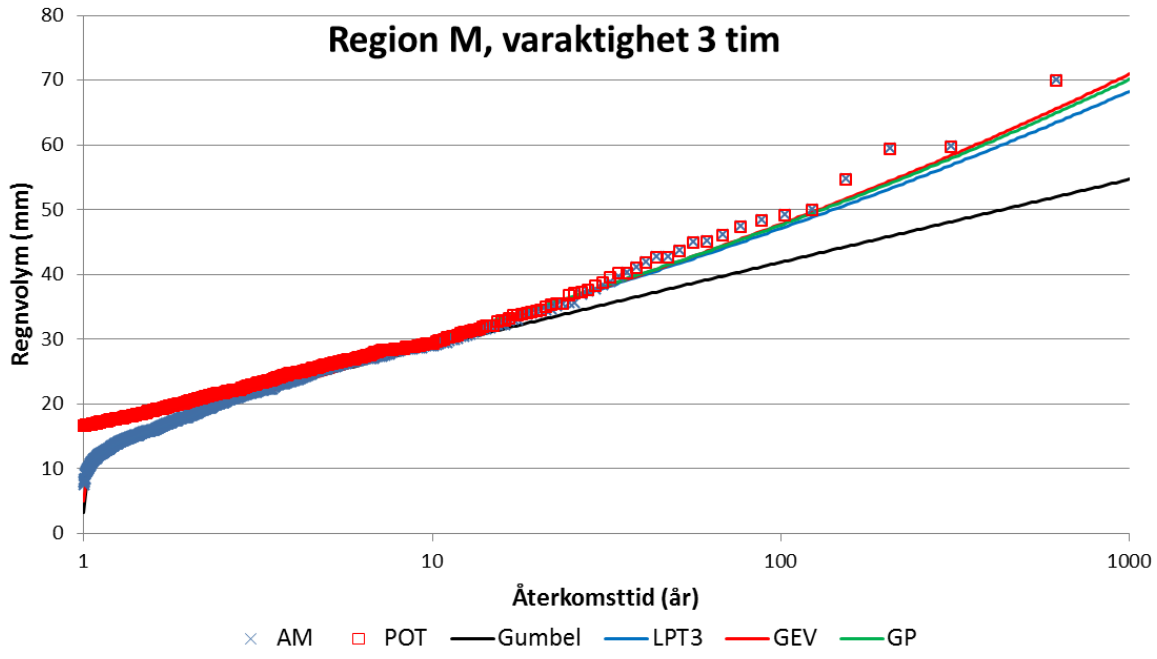


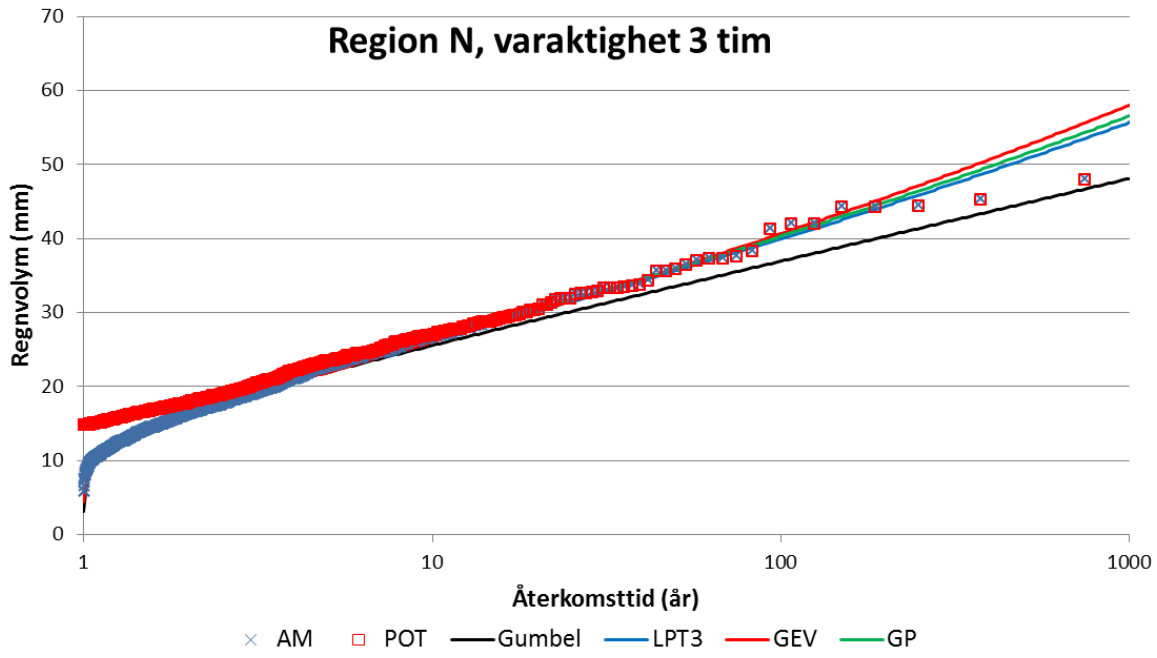
Figur 1. (forts.) Framtagna 15-min extremvärden och anpassade sannolikhetsfördelningar för de sammanslagna serierna i region M och N.





Figur 2. Framtagna 3-tim extremvärden och anpassade sannolikhetsfördelningar för de sammanslagna serierna i region SV och SÖ.

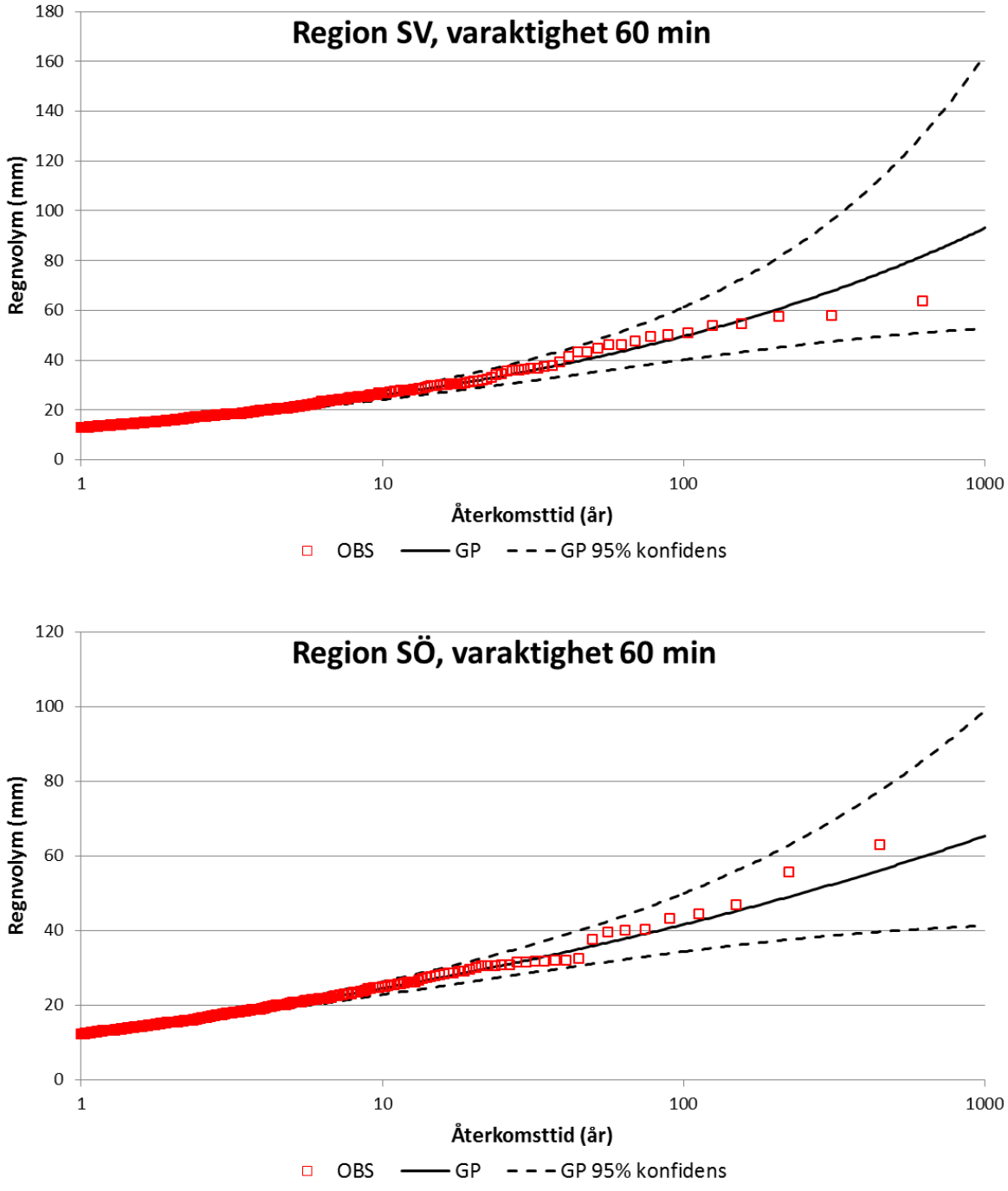




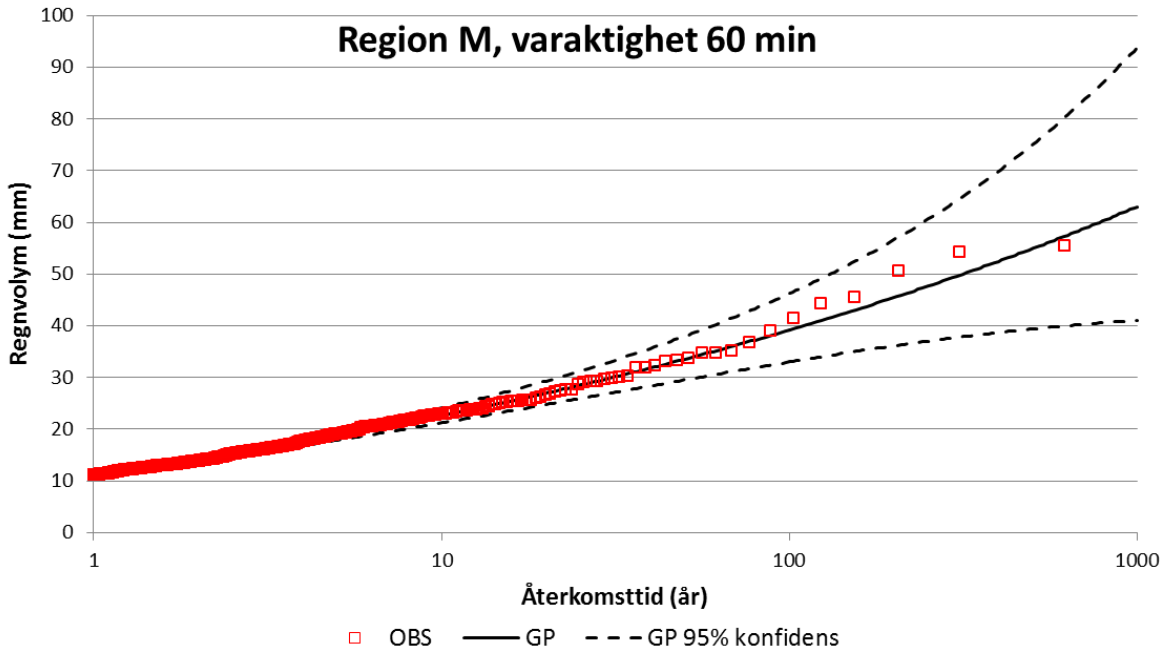
Figur 2. (forts.) Framtagna 3-tim extremvärden och anpassade sannolikhetsfördelningar för de sammanslagna serierna i region M och N.

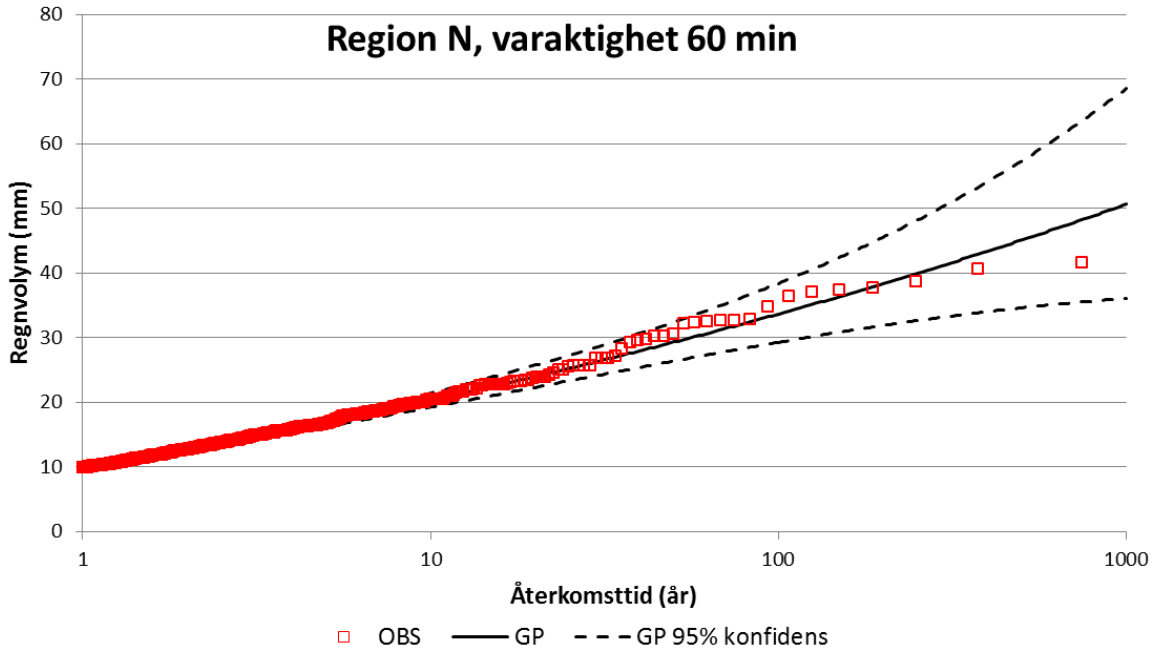
Bilaga II.5 Modellering av konfidensintervall

Såsom beskrivs i avsnitt 3.1.2.2 valdes POT-metoden med GP-fördelningen ut för att ta fram den slutliga statistiken. För att kvantifiera osäkerheten i anpassningarna beräknades konfidensintervall. Olika konfidensgrader testades och till slut valdes 95%, som kan anses vara standard. Figur 3 visar konfidensintervallen för anpassningen till 60-min värden från samtliga regioner.



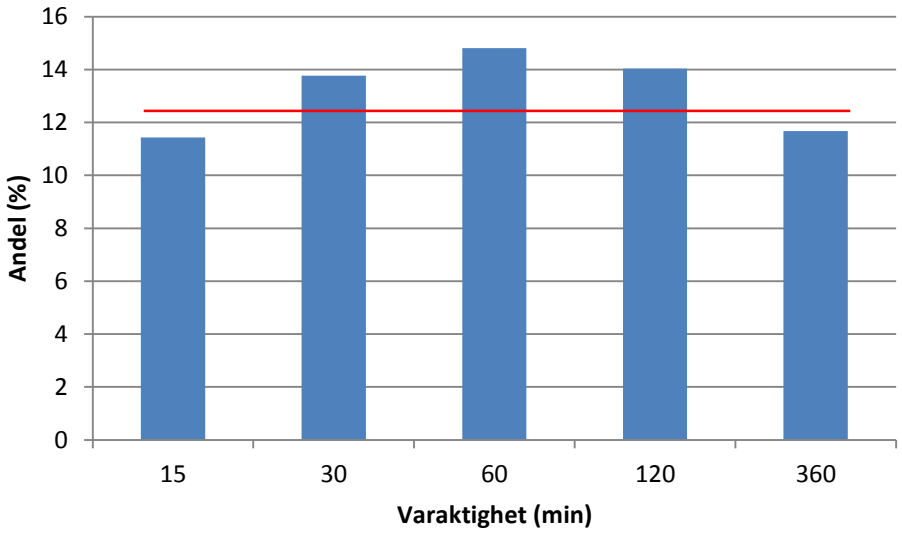
Figur 3. Framtagna 1-tim extremvärden och anpassad GP-fördelning inklusive konfidensintervall för de sammanslagna serierna i region SV och SÖ.





Figur 3. (forts.) Framtagna 1-tim extremvärden och anpassad GP-fördelning inklusive konfidensintervall för de sammanslagna serierna i region M och N.

I Figur 3 ses att konfidensintervallet är nära symmetriskt upp till minst 100 års återkomsttid (i själva verket är det övre intervallet marginellt större än det undre). Därför antogs en modell för att uttrycka intervallet som ± en andel (%) av själva ackumulationen för denna varaktighet. Denna andel visade sig ha en måttlig variation över olika varaktigheter, se ett exempel i Figur 4, och antogs därför kunna beskrivas av medelvärdet över alla varaktigheter.

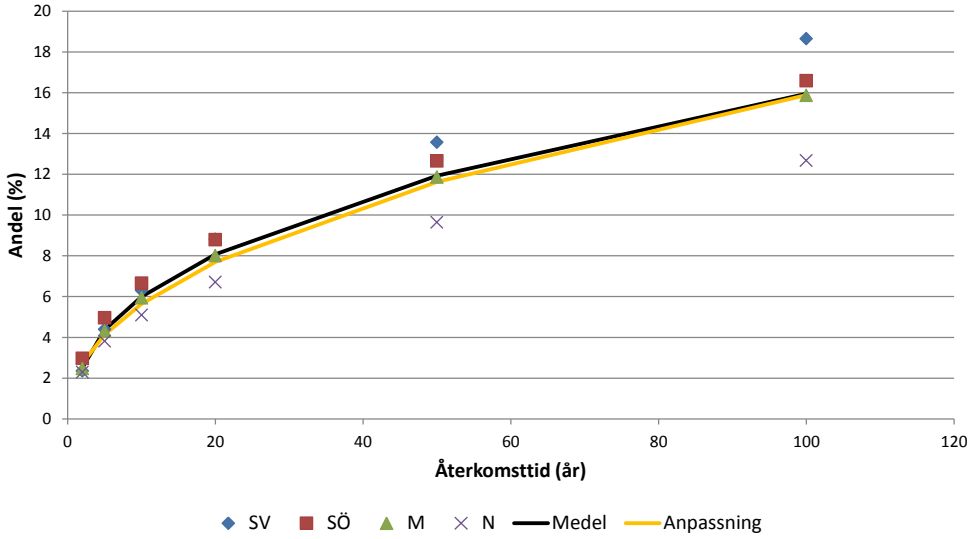


Figur 4. Konfidensintervallets andel av ackumulationen för olika varaktigheter med 50 års återkomsttid i region SÖ. Röd linje visar medelvärdet.

Andelen varierade däremot tydligt med återkomsttid, från ett fåtal procent för korta återkomsttider upp till 15-20% för långa. En enkel potensfunktion användes för att beskriva denna variation, för alla varaktigheter

$$Andel = 2 \times Återkomsttid^{0.45}$$

vilken väl beskriver medelkurvan (Figur 5). Som synes finns viss regional variation, med störst andel i region SV och lägst i region N. Skillnaderna på ett fåtal procent påverkar emellertid väldigt lite de slutliga konfidensintervallen; därför försummas den regionala variationen och samma anpassning används för alla regioner.



Figur 5. Konfidensintervallets andel av ackumulationen för olika återkomsttider i de olika regionerna, för medelvärdet av dem samt som en anpassad funktion.

Bilaga II.6 Metod för beräkning av extremvärdesstatistik i griddade modeller

En peak-over-threshold-metod (POT) används för extremvärdesanalys i de griddade data. För varje gridpunkt och varaktighet görs följande:

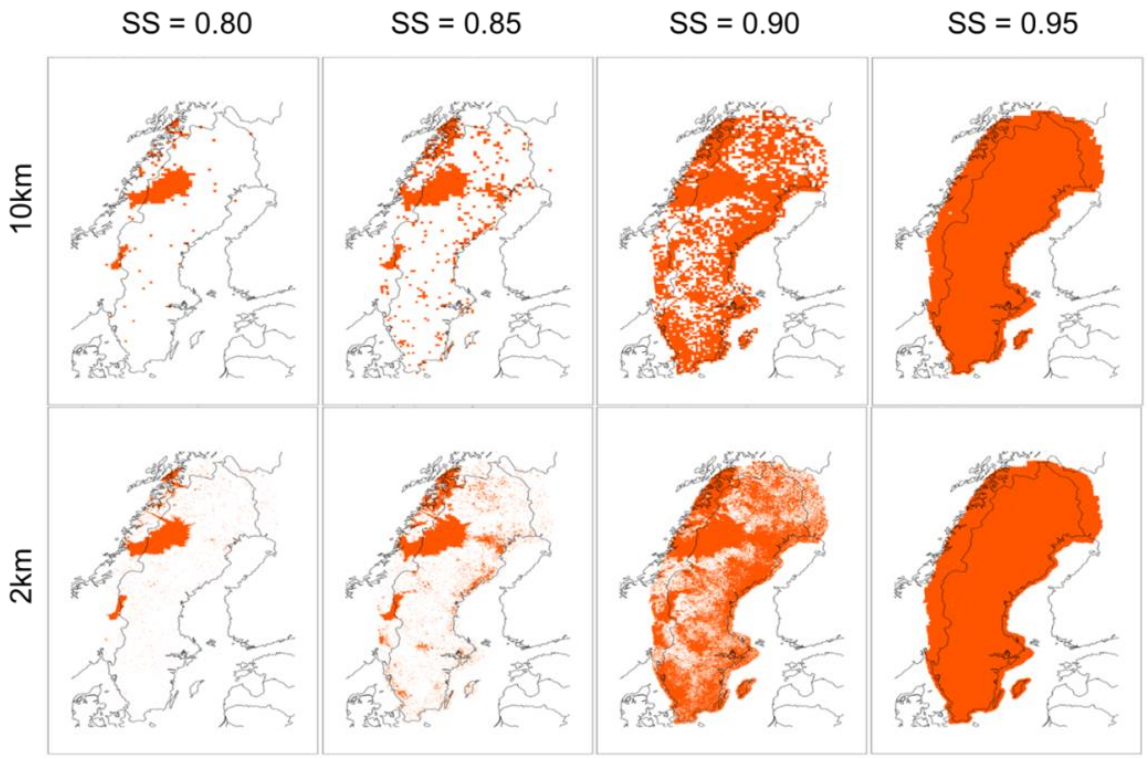
1. För en given period av 15 till 30 år (beroende på seriens längd) sorteras nederbördsmängden i avtagande ordning.
2. Den högsta nederbördshändelsen noteras
3. *X* tidssteg före och efter händelsen tas bort från analysen på grund av att de inte kan anses oberoende samplingar. För varaktigheter på en till tre timmar är *X* tre timmar och för längre varaktigheter sätts *X* lika med varaktigheten.
4. Upprepa över steg 2-3 tills 650 händelser har registrerats över 30 år, eller 325 händelser över 15 år för HIPRAD.
5. Anpassa en Generalized Pareto (GP) fördelning till data och beräkna återkomsttider.

Statistik för de fyra regionerna beräknas sedan genom medelvärdesbildning av resultatet i vardera gridpunkt.

Gränsen 650 värden, det vill säga i snitt ca 22 värden per år, togs fram genom att studera parametervärdena för GP vid olika tröskelvärden. Parametervärdena var stabila inom brusnivån fram till ungefär 650 datapunkter varefter fördelningen ändrar form. Det kan tolkas som att det inte längre är svansen på fördelningen som samplas och att extremvärdesteorin inte längre gäller. Det är en fördel att ligga så nära den punkten som möjligt för att få robusta resultat. Utvärderingen utfördes på punkter i Sverige, men även runt om i Europa på klimatmodellerna.

Bilaga II.7 Utsortering av orimliga gridpunkter i HIPRAD

Metoden för att sortera ut orimliga data från HIPRAD består i att undersöka den normaliserade intensitetsfördelningen (alltså en sannolikhetsfördelning) i varje gridpunkt jämfört med medelintensitetsfördelningen för de tjugo närmaste automatstationerna. För att utvärdera fördelningen användes en metod där man summerar den överlappande arean av två sannolikhetsfördelningar, vilket ger ett värde (SS) mellan noll och ett, där ett innebär en exakt likhet. Statistiskt brus gör att en perfekt överensstämmelse inte är rimlig, så flera olika nivåer undersöktes. Figur 6 visar resultatet för HIPRAD på grundupplösningen samt på upplösningen för de regionala klimatmodellerna. Orimligt många punkter ratas vid gränsen 0.9, så den något lägre gränsen 0.85, det vill säga 85% överensstämmelse med stationer, antogs för att sortera bort vissa gridpunkter. Vissa av regionerna som visas i figuren för SS=0.85 känns igen som problemregioner för radarn medan andra eventuellt faller bort på grund av den exakta gräns som satts även om de inte avviker drastiskt.



Figur 6. Resultat av utsortering av HIPRAD-pixlar för olika gränsvärden och för HIPRAD på två olika upplösningar. I orange visas de gridpunkter som har lägre SS-värden än titeln anger.