CHALMERS

GÖTEBORGS UNIVERSITET

Analys av trender i extrema vattenflöden i Sverige

Assessment of trends in extreme waterflows in Sweden

Kandidatarbete inom civilingenjörsutbildningen vid Chalmers

Edvin Ahlström
Nils Hammar
Vincent Harbander
Simon Rödén
Matilda Savolainen

Institutionen för Matematiska vetenskaper
CHALMERS TEKNISKA HÖGSKOLA
GÖTEBORGS UNIVERSITET
Göteborg, Sverige 2024

Analys av trender i extrema vattenflöden i Sverige

Kandidatarbete i matematik inom civilingenjörsprogrammet Teknisk fysik vid Chalmers
Edvin Ahlström

Kandidatarbete i matematik inom civilingenjörsprogrammet Teknisk matematik vid Chalmers
Vincent Harbander
Simon Rödén

Kandidatarbete i matematik inom civilingenjörsprogrammet Bioteknik vid Chalmers
Nils Hammar

Kandidatarbete i matematik inom civilingenjörsprogrammet Samhällsbyggnadsteknik vid Chalmers
Matilda Savolainen

Handledare: Holger Rootzén

Institutionen för Matematiska vetenskaper
CHALMERS TEKNISKA HÖGSKOLA
GÖTEBORGS UNIVERSITET
Göteborg, Sverige 2024

Förord

Vi vill tacka vår handledare Holger Rootzén för hans entusiasm för projektet och professionella hjälp. Gruppen tackar Stuart Coles, vars lärobok *An Introduction to Statistical Modeling of Extreme Values*, ligger till grund för en stor del av teorin i rapporten. Dessutom tackas SMHI för deras lättillgängliga data och hjälpsamma korrespondens.

Under arbetsprocessen har individuella prestationer dokumenterats i en individuell loggbok. Veckovis har denna loggbok samt mötesanteckningar sammanfattats i en dagbok som har skrivits i ett roterande schema. Varken loggboken eller dagboken är bilagd i rapporten. Vi är som grupp eniga att alla har gjort sin del i projektet. Nedan presenteras en bidragsrapport som klargör ansvarsfördelningen för den skriftliga rapporten.

Bidragsrapport		
-	Rubrik	Författare
-	Populärvetenskaplig presentation	Matilda & Vincent
-	Sammandrag	Matilda & Vincent & Simon
-	Abstract	Vincent & Simon & Matilda
1	Inledning	Matilda & Edvin
1.1	Bakgrund	Matilda & Vincent & Edvin
1.2	Syfte och frågeställningar	Samtliga
1.3	Avgränsningar	Samtliga
2	Teori	Samtliga
2.1	Extremvärdesanalys	Nils & Simon & Vincent & Matilda
2.1.1	Generaliserade extremvärdesfördelningen	Vincent
2.1.2	Block maxima-metoden	Matilda & Vincent
2.1.3	Generaliserade Paretofördelningen	Vincent
2.1.4	Tröskelmetoden	Vincent & Edvin
2.2	Modeller med trend	Edvin & Vincent
2.3	Trend i frekvensen λ i Poissonprocessen	Vincent
2.4	Maximum likelihood-skattning	Nils & Vincent
2.5	Hypoteser och tester	Vincent
2.5.1	Test av nollskillda parametrar	Vincent
2.5.2	Anderson-Darlingtest	Simon
2.5.3	Benjamin-Hockbergproceduren	Vincent
2.6	Återkomstnivåplott	Matilda & Vincent
2.7	Icke-stationära alternativ till återkomstnivåplott	Vincent
2.7.1	Projektionsnivåplott	Vincent
2.7.2	Prediktionsnivåplott	Vincent
3	Metod	Vincent
3.1	Modeller	Simon
3.2	Datainsamling och databehandling	Edvin & Nils
3.3	Dataanalys	Simon & Matilda & Vincent
3.4	Monte Carlo-sampling	Vincent
3.5	Riskförändring	Vincent & Edvin
4	Resultat	Vincent
4.1	Modellers giltighet	Edvin & Matilda & Nils
4.2	Trender i modeller	Edvin & Matilda & Nils
4.3	Vädervarningar	Edvin & Matilda
5	Diskussion	Samtliga
5.1	Samhälleliga och etiska aspekter	Matilda & Simon & Vincent
5.2	Slutsatser	Samtliga

-	Appendix A.1	Vincent
-	Appendix A.2	Vincent
-	Appendix A.3	Simon
-	Appendix A.4	Vincent
-	Appendix B.1	Edvin
-	Appendix B.2	Edvin
-	Appendix B.3	Simon
-	Figurer	Edvin
-	Tabeller	Edvin
-	LaTeX-generator för bilagan	Vincent
-	Monte Carlo-samplern	Vincent
-	Dataanalys	Simon
-	Exempel av plottgenerering i Python	Edvin
-	Korrekturläsning	Samtliga

Populärvetenskaplig presentation

Kommer klimatförändringar leda till fler eller större extrema vattenflöden i Sverige? Detta skulle i så fall resultera i ökade risker för kostsamma och eventuellt farliga översvämningar, vilket kräver omfattade investeringar i nya skyddsåtgärder. Med vetskap om dessa risker har förståelse för och hantering av extrema vattenflöden blivit alltmer akut.

För att uppskatta dessa trender kan matematiska modeller användas. Inom statistik benämns studier av extrema händelser extremvärdesanalys. De vanligaste modellerna för att studera extrema händelser är block maxima-metoden och tröskelmetoden. Block maxima-metoden delar upp datan i tidsperioder, exempelvis år, och analyserar den största händelsen i varje tidsperiod. Tillämpat till årliga vattenflöden hade block maxima endast analyserat det största vattenflödet varje år för att förutse stora vattenflöden i framtiden. Alternativt kan tröskelmetoden användas. Metoden bygger på att välja ett tröskelvärde och därefter analysera händelser som är större än tröskeln. Tröskeln väljs tillräckligt högt för att endast en låg andel av datan ska överskrida den, exempelvis 2 % av mätpunkterna. Därefter analyseras alla mätpunkter över tröskelvärdet för att förutse förekomsten av extrema händelser.

Målet med projektet är att undersöka om det finns trender i frekvensen och storleken av extrema vattenflöden i Sverige samt hur dessa eventuella trender ser ut. Orsakerna bakom eller potentiella skador kopplade till de eventuella trenderna kommer inte att undersökas.

De statistiska modellerna för trender bygger på fördelningar där vi antar att vissa parametrar har ett tidsberoende. För att analysera trenderna användes 5 olika modeller. Block maxima-modellen utan trender analyserar sannolikheten för extrema händelser utan att ta hänsyn till förändringar över tid. Block maxima-modellen med trender i läge antar att lägesparametern i fördelningen för block maxima förändras linjärt med tiden. Om lägesparametern ökar över tid innebär det att de extrema händelserna blir större. Nästa modell är block maxima med exponentiell trend i skala. Om trenden är positiv innebär det att extrema händelser blir mer volatila. Tröskelmetoden utan trender undersöker händelser som överstiger ett förbestämt tröskelvärde. Tröskelmetoden med trend i skala kan precis som block maxima-modellen användas för att avgöra om extrema händelser blir mer eller mindre volatila över tid.

Parametrarnas värde kan vara svårtolkade. Därför användes SMHIs vädervarningar för att presentera resultatet. Vädervarningarna är anpassade för att kommunicera risker av översvämning. Vi jämför förhållandena mellan risken idag och risken om 30 år. Ungefär en femtedel av vattendragen hade en statistiskt signifikant trend i volatilitet, som var negativa i de flesta fallen. I 7 % av vattendragen observerades en signifikant trend i frekvensen av extrema vattenflöden.

Sammandrag

Denna rapport har utforskat trender i extrema vattenflöden i 54 svenska vattendrag med hjälp av extremvärdesteori och data från SMHI. Syftet var att analysera eventuella förändringar i frekvens och storlek av extrema vattenflöden över tid. För att uppnå detta användes GEV- och GP-fördelningar ifrån block maxima-metoden respektive tröskelmetoden. GEV- och GP-fördelningarna ansattes med olika trenderparametrar och alla parametrar maximum likelihood-skattades. För att avgöra om modellerna utan trender var lämpliga användes Anderson-Darlingtest. Dessutom utfördes tester för att ta reda på om trendparametrarna var statistiskt signifikant skilda från noll. Därefter justerades p-värden från testerna med Benjamini-Hochberg individuellt per modell och test. För att underlätta förståelsen för resultatet redovisas riskförändingsfaktorer byggda på SMHIs vädervarningar. Riskerna uppskattas genom Monte Carlo-sampling. Två nya plottar för att redovisa extremvärdesrisker presenteras även. Resultatet visar att det inte finns vidsträckta trender i någon modell, men att det finns statistiskt signifikanta trender för block maxima i skalparametern i 10 av de 54 vattendrag som undersöktes. Av dessa var 8 av 10 negativa trender. Vidsträckta trender i frekvensen av extrema vattenflöden hittades inte heller. Tröskelmetoden visade sig passa dåligt för datan och därför dras inga slutsatser ifrån den. En bilaga publiceras med fullständiga resultat för alla vattendrag och modeller.

Abstract

This report has assessed the existence and character of trends in extreme water flows in 54 swedish streams using extreme value analysis and data gathered by SMHI, the Swedish Meteorological and Hydrological Institute. The aim was to analyze possible changes in frequency and magnitude of extreme water flows over time. The methods block maxima and peaks over threshold were used with their respective distributions GEV and GP to achieve this end. The GEV and GP distributions were used with different trend parameters and all parameters were maximum likelihood estimated. The Anderson-Darling test was used to determine if the models without trends were appropriate. Statistical tests were used to determine if the trend parameters and the shape parameter were statistically significantly different from 0. Then the p-values were adjusted using the Benjamini-Hochberg procedure individually per model and test. SMHI's weather warnings and risk ratios were used to increase the accessibility of the results. Monte Carlo sampling was used to estimate risks. Additionally, two novel extreme value risk plots are introduced. Our results show no statistically signficant widespread trends in any models, but there were significant trends in 10 out of 54 streams for the block maxima model with trend in the scale parameter. 8 out of 10 of these had negative trends. Widespread trends in the frequency of extreme water flows was not found either. The peaks over threshold model fit the data poorly, so no conclusions were drawn regarding it. A supplemental document was also produced containing the complete results for all streams and models.

Innehåll

1 Inledning

Genom alla tider har människor varit beroende av vattendrag för olika ändamål, från livsmedelsproduktion till transportnät och energikällor. Därför har samhällen ofta etablerat sig längs med naturliga vattendrag. Närheten till vattendrag medför dock översvämningsrisker [1] som kan leda till allvarliga ekonomiska och ekologiska skador samt innebära livsfara. Det är därför viktigt att undersöka beteendet och eventuella trender av extrema vattenflöden. Genom att utveckla och tillämpa modeller för att bedöma översvämningsrisker blir det möjligt att rationellt prioritera skyddsåtgärder och infrastruktur som vallar, diken och kraftverk för att minimera skador. Francis m.fl. [2] har undersökt översvämningar, tropiska stormar, torka och jordbävningar på hela jordklotet och fann att frekvensen för samtliga fenomen ökade under perioderna 1963-1967 och 1988-1992. Väderrelaterade naturkatastrofer visade en särskilt stor ökning.

Högre globala temperaturer förväntas enligt modeller leda till en ökning av mängden vatten som cirkulerar genom vattnets kretslopp. Detta beror på att högre temperatur leder till ökad avdunstning och en större kapacitet för luften att hålla fuktigthet. Högre temperaturer kan även öka konvektionen över varmare länder och havsytor vilket leder till en mindre stabil atmosfär och ytterligare ökar risken för kraftig nederbörd. Enligt klimatmodeller förväntas denna ökning av nederbörd fördelas ojämnt över världen [2].

1.1 Bakgrund

I projektet analyseras vattenflöde, även kallat vattenföring, i ett antal svenska vattendrag. Vattenflöde är den volym vatten som rör sig genom ett vattendrag vid en given plats under en viss tid [3]. Analysen av vattenflöden bygger på data insamlad av SMHI, Sveriges Meteorogiska och Hydrologiska Insititut. SMHI:s data för vattenflöden beräknas vid majoriteten av mätstationerna utifrån sambandet mellan vattennivå och vattenflöde via en så kallad avbördningskurva som är specifik för varje station [4]. Vetenskapen som studerar vattenflöde kallas hydrologi. U.S. Geological Survey, en amerikansk statlig organisation, använder följande definition för hydrologi [5].

> Hydrologi är vetenskapen som omfattar förekomsten, fördelningen, rörelsen och egenskaperna av vattnet på jorden och dess förhållande till omgivningen inom varje fas av den hydrologiska cykeln. (Citatet är översatt av oss.)

Hydrologer delar oftast inte upp mätdata utifrån kalenderår utan istället delas tiden in i så kallade vattenår. Olika organisationers definitioner för vattenårets start- och sluttider varierar något. Detta är för att kunna förklara vattenflöden i ett vattenår med nederbörd i samma vattenår. Därför börjar och slutar vattenåret i en torr säsong av året [6]. U.S Geological Survey definierar det nuvarande vattenårets slut som den 30:e september och nästa års början till den 1:a oktober [7], vilket är definitionen som används i rapporten.

För att förstå den potentiella påverkan av vattenflöden och andra väderfenomen har SMHI definierat vädervarningar. Vädervarningarna ges på nivåerna gul, orange eller röd, utifrån hur stora konsekvenser vädret kan medföra. Röd varning är den allvarligaste, men alla varningsnivåer signalerar potentiellt problematiskt väder [8]. Varningarna bygger på återkomsttid, vilket är ett begrepp för att kvantifiera hur ofta en viss händelse uppstår. I genomsnitt förväntas händelsen ske en gång under händelsens återkomsttid. Gul varning innebär ett vattenflöde med återkomsttid 5-25 år, orange varning 25-50 år och röd varning över 50 år [9].

Vädervarningssystemet är en skala för att bedöma potentiella samhällskonsekvenser och risker för skador på egendom och miljö. Gul varning indikerar risk för möjliga störningar i samhällstjänster, som kollektivtrafik, där vissa områden kan vara mer utsatta än andra. Vid orange varning förväntas allvarligare konsekvenser, inklusive risk för betydande skador och mer utbredda störningar i samhällstjänster. Röd varning indikerar en akut situation med stor fara för allmänheten, en hög sannolikhet för mycket allvarliga skador på egendom och miljö samt betydande störningar i kritiska samhällstjänster som kollektivtrafik och räddningstjänst [8]. Varningarna, förutom att redovisas på SMHI:s hemsida och väderapp, skickas direkt till länsstyrelser, kraft- och vattenregleringsföretag, vissa centrala och regionala myndigheter samt massmedia. Varningarna är viktiga beslutsunderlag och kan exempelvis leda till omställning av kollektivtrafik samt förberedelser för insatser på

vägnätet [9].

Tidigare forskning har visat att översvämningar i vissa europeiska floder troligtvis har påverkats av klimatförändringar medan andra vattendrag visar en mindre tydlig koppling [10]. Även SMHI har undersökt vattendrag och fann då en ökande trend i extrema vattenflöden i vissa delar av Sverige, men en minskande trend i andra delar [11]. Mänsklig infrastruktur påverkar också översvämningar och kan leda till en ökning av de högsta vattenflödena. Det innebär en ökning av storleken samt frekvensen av översvämningar [12].

För att analysera trender hos ovanliga händelser som extrema vattenflöden kan både block maxima-metoden [13, s. 45-69] och tröskelmetoden, även kallad *peaks over threshold* [13, s. 74-86], från extremvärdesteori användas. Block maxima-metoden analyserar maxvärden inom stora tidsperioder, kallade *block*, och används för att förutsäga extrema händelser inom en lång tidsperiod. Maxima av dagliga genomsnitt av vattenflöden i ett vattendrag över ett år, eller en annan lång tidsperiod, kan förväntas approximativt tillhöra en känd fördelningsfamilj [13, s. 49].

Tröskelmetoden använder däremot händelser som överstiger ett förbestämt tröskelvärde för att bedöma risker för framtida extrema händelser. Tröskelmetoden förutsätter att händelserna är oberoende och likfördelade. Ett tröskelvärde väljs och endast tidpunkterna då flödet överstiger detta tröskelvärde undersöks [13, s. 74]. Flödena som överstiger tröskelvärdet följer approximativt en annan känd fördelningsfamilj för stora tröskelvärden [13, s. 75-77]. Metoderna från extremvärdesteori ger oss skattningar av sannolikheter för extrema vattenflöden imorgon eller inom en tidsperiod. Utifrån dessa kan frekvensen och storleken av översvämningar undersökas.

1.2 Syfte och frågeställningar

Projektets syfte är att undersöka om det finns trender i frekvensen och storleken av extrema vattenflöden i Sverige samt hur dessa eventuella trender ser ut. Följande frågeställningar kommer att besvaras i projektet.

- Passar block maxima- och tröskelmetoden för att analysera extrema vattenflöden i Sverige?
- Existerar det trender i extrema vattenflöden i Sverige och vad är i så fall den förändrade risken för vädervarningarna?

1.3 Avgränsningar

Endast data mellan 1960-2022 kommer användas i analysen. Dataserier som saknar en eller flera datapunkter inom det relevanta tidsspannet analyseras inte. Endast allmänt tillgänglig data från SMHI kommer att användas. Datan kommer inte kvalitetstestas och eventuella förändringar av vattendrag eller mätmetoder kommer inte tas hänsyn till. För block maxima-metoden kommer blocken bestå av vattenår med start den 1:a oktober och slut den 30:e september. Det kommer inte att undersökas ifall ett annat val av blockindelning hade påverkat resultaten. Orsakerna bakom eller potentiella skador från de extrema vattenflödena kommer inte heller att analyseras. En mindre ökning i en tätort kan exempelvis ha större konsekvenser än en översvämning i ett glesbefolkat område. Hur eventuella trender i de undersökta vattendragen kan användas för att förutsäga trender hos andra vattendrag kommer inte heller bedömas.

2 Teori

I detta avsnitt läggs grunden för metoderna som tillämpas i projektet. Med en utgångspunkt i extremvärdesanalys diskuteras skattningar, anpassningar av de så kallade GEV- och GP-fördelningarna för trender samt hypotestestning. Användning av Monte Carlo-simulering för att generera alternativ till återkomstnivåplottar för modeller med trender presenteras också.

2.1 Extremvärdesanalys

Stuart Coles [13, s. 1-2] definierar extremvärdesanalys på följande sätt.

> Den utmärkande funktionen av extremvärdesanalys är målet att kvantifiera det stokastiska beteendet av en process på en ovanligt stor - eller liten - nivå. I extremvärdesanalyser efterfrågas oftast skattningar för sannolikheter för händelser som är mer extrema än något tidigare observerat. Antag för sakens skull att en vall som översvämningsskydd har kravet att det måste skydda mot alla havsnivåer som den troligen kommer stöta på under sin beräknade livstid på, säg 100 år. Lokal data om havsnivåer kanske finns, men bara för en mycket kortare tidsperiod, säg 10 år. Utmaningen är då att uppskatta vilka havsnivåer som kan inträffa de nästa 100 åren, givet 10-årshistoriken. Extremvärdesteori ger ett ramverk som tillåter sådana extrapolationer. (Citatet översatt av oss.)

Antag till att börja med att $X_1, \ldots, X_n$ är likfördelade slumpvariabler med fördelningsfunktion $F(x)$. Vi definierar

$$M_n = \max\{X_1, ..., X_n\}. \tag{1}$$

Om tidsmellanrummet är ett dygn och $n = 365$ skulle M_n vara årsmaximat. Fördelningsfunktionen M_n relaterar till fördelningsfunktionerna F för X_i genom

$$\begin{aligned} \mathrm{P}\left(M_n \leq x\right) &= \mathrm{P}\left(X_1 \leq x, \ldots, X_n \leq x\right) \\ &= \mathrm{P}\left(X_1 \leq x\right), \ldots, \mathrm{P}\left(X_n \leq x\right) \\ &= \left(F(x)\right)^n \end{aligned} \tag{2}$$

[13, s. 45].

Eftersom F är okänd och små avvikelser i F leder till stora avvikelser i F^n är det svårt att bestämma F. Istället efterliknar extremvärdesstatistikens tillvägagångssätt den klassiska statistiken [13, s. 45-46]. I klassisk statistik används ofta den centrala gränsvärdessatsen, vilket säger att medelvärden av oberoende, likfördelade slumpvariabler med ändligt väntevärde och varians är asymptotiskt normalfördelade [14, s. 268]. Likt den centrala gränsvärdessatsen finns ett asymptotiskt resultat för maximum av oberoende slumpvariabler som extremvärdesstatistiken utnyttjar.

Om det existerar en talföljd (b_n) och en positiv talföljd (a_n) så att

$$\mathrm{P}\left(\frac{M_n - b_n}{a_n} \leq x\right) \xrightarrow[n \longrightarrow \infty]{} G(x), \tag{3}$$

där G är en fördelningsfunktion för en icke-degenererad fördelning, måste G vara en GEV-fördelningsfunktion, som introduceras i nästa avsnitt [13, s. 46-47]. En degenererad fördelning är en sannolikhetsfördelning som har ett specifikt värde med sannolikhet 1.

2.1.1 Generaliserade extremvärdesfördelningen

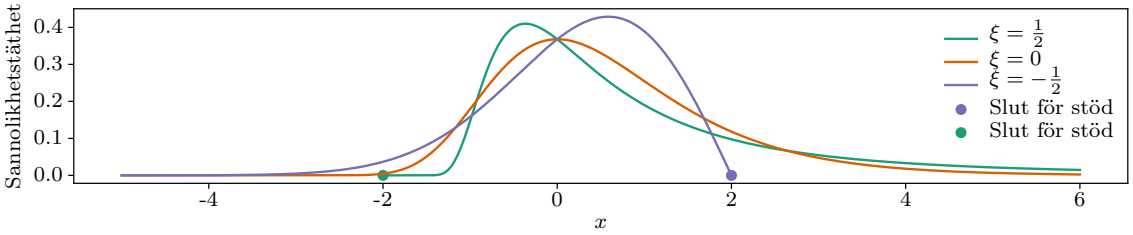
Generaliserade extremvärdesfördelningen, ofta förkortad GEV-fördelningen, är en familj fördelningar som bestäms av tre parametrar. Dessa är parametern $\mu \in \mathbb{R}$ kallad läge, parametern $\sigma > 0$

kallad skala samt parametern $\xi \in \mathbb{R}$ som kallas form. En anmärkning är att μ och σ inte är väntevärdet och standardavvikelsen för fördelningen. GEV-fördelningarna är unimodala och består av tre familjer fördelningar med olika stöd, vilket är området där täthetsfunktionen är skild från 0. För $\xi < 0$, $\xi = 0$ och $\xi > 0$ blir GEV en omvänd Weibullfördelning, Gumbelfördelning respektive Fréchetfördelning. Se figur 1.

GEV-fördelningsfunktionen är

$$F(x) = \begin{cases} \exp\left(-e^{-\frac{x-\mu}{\sigma}}\right), & \xi = 0 \\ \exp\left(-\left(1+\xi\left(\frac{x-\mu}{\sigma}\right)\right)^{-\frac{1}{\xi}}\right), & \xi \neq 0 \end{cases} \tag{4}$$

och har stöd på $\left\{x \in \mathbb{R} \;\middle|\; 1+\xi\left(\frac{x-\mu}{\sigma}\right) > 0\right\}$. Stödet blir därför $\mathbb{R}$ för $\xi = 0$, $\left[\mu - \frac{\sigma}{\xi}, \infty\right)$ för $\xi > 0$ och $\left(-\infty, \mu - \frac{\sigma}{\xi}\right]$ för $\xi < 0$ [13, s. 47-48].



Figur 1: Täthetsfunktion för GEV-fördelningen med olika ξ-värden. Här är $\mu = 0$ och $\sigma = 1$.

Observera att Gumbelfördelningsfunktionen uppstår som ett gränsvärde av GEV-fördelningsfunktionen när $\xi \longrightarrow 0$. För positiva ξ får GEV-fördelningarna feta svansar. För $\xi \geq \frac{1}{2}$ är variansen av GEV-fördelningen oändlig och för $\xi \geq 1$ är väntevärdet oändligt.

Av samband 3 följer det att maximum av oberoende likfördelade GEV-fördelningar också är GEV-fördelad. Låt $X_1, \ldots, X_n \sim \text{GEV}(\mu, \sigma, \xi)$ vara obereoende slumpvariabler. Då är

$$\max\{X_1, \ldots, X_n\} \sim \text{GEV}(\mu^*, \sigma^*, \xi^*), \tag{5}$$

där

$$\begin{aligned} \mu^* &= \mu + \frac{\sigma}{\xi}(n^\xi - 1), \\ \sigma^* &= \sigma n^\xi, \\ \xi^* &= \xi \end{aligned} \tag{6}$$

om $\xi \neq 0$. Annars är

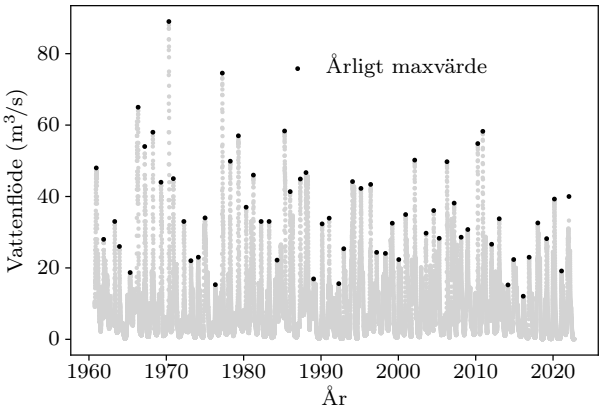
$$\begin{aligned} \mu^* &= \mu + \sigma \ln n, \\ \sigma^* &= \sigma, \\ \xi^* &= \xi. \end{aligned} \tag{7}$$

Se appendix A.1 för beviset.

2.1.2 Block maxima-metoden

Block maxima är en metod från extremvärdesstatistik som utnyttjar samband 3 för att bedöma risker av extrema händelser. Antag att det finns mätpunkter som observerats vid olika tidpunkter. Datan delas först upp i tidsintervaller, så kallade *block*, där intervallängden ofta bestäms utifrån ett praktiskt perspektiv, vilket vanligtvis leder till årliga, månatliga eller dagliga blocklängder. Inom varje block identifieras det maximala värdet. Se figur 2. Vid tillräckligt stora block är dessa maxvärden approximativt GEV-fördelade på grund av samband 3. Större block ger färre maxima och större varians i uppskattningen medan mindre block troligtvis inte passar GEV-fördelningen lika väl och kan leda till bias [15].

Maxvärden av oberoende eller lokalt beroende mätvärden är oberoende i gräns när blocklängden går mot oändligheten. Om mätvärdena dessutom antas vara likfördelade är maxvärdena oberoende stickprov av en GEV-fördelning. Detta möjliggör skattning av dess parametrar [13, s. 92-93]. Genom att skatta parametrarna kan sannolikheter för maximala mätvärden de kommande åren uppskattas.



Figur 2: Exempel för val av datapunkter i block-maxima modellen ifrån vattendraget Getebro.

2.1.3 Generaliserade Paretofördelningen

Generaliserade Paretofördelningen, ofta förkortat GP-fördelningen, tillåter ett annat tillvägagångssätt till extremvärdesteorin. Låt $X_1, X_2, X_3, \ldots$ vara oberoende, likfördelade slumpvariabler med fördelningfunktion F och beteckna maximum av de första n slumpvariablerna $M_n = \max\{X_1, \ldots, X_n\}$. Antag att M_n uppfyller samband 3. Definiera

$$F_u(x) = \mathrm{P}\left(X_i - u < x | X_i > u\right) \tag{8}$$

för något X_i. Vilken av dem är inte relevant på grund av deras likfördelning. Funktionen F_u är då fördelningsfunktionen för $X_i - u | X_i > u$, alltså X_i:s överstigning över u, givet att den överstiger. GP-fördelningen motiveras av satsen nedan. Antag att det existerar en funktion $b(u)$ och en positiv funktion $a(u)$ så att

$$F_u\left(\frac{x - b(u)}{a(u)}\right) \xrightarrow[u \longrightarrow \infty]{} G(x), \tag{9}$$

för någon icke-degenererad slumpvariabels fördelningsfunktion G. Då är G en fördelningsfunktion för en GP-fördelning [16, s. 793-798]. Svansen av X_i är GP-fördelade om och endast om den uppfyller samband 3 [17, s. 125].

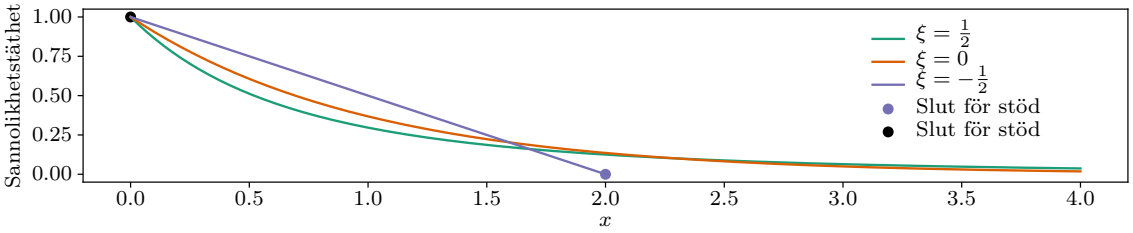
En informell följdsats av samband 9 är att om $M_n \approx \mathrm{GEV}(\mu, \sigma, \xi)$ för stora n, där $\mu \in \mathbb{R}$, $\sigma > 0$ och $\xi \in \mathbb{R}$, så gäller

$$X - u \mid X > u \approx \mathrm{GP}(\tilde{\sigma}, \xi) \tag{10}$$

för $\tilde{\sigma} = \sigma + \xi(u - \mu)$ [13, s. 75]. Detta är hur den används i praktiken. GP-fördelningen beskriver alltså svansen av många fördelningar. Fördelningen har två parametrar, $\sigma > 0$ kallad skala, och $\xi \in \mathbb{R}$ kallad form. Fördelningsfunktionen är

$$F(x) = \begin{cases} 1 - \left(1 + \frac{\xi x}{\sigma}\right)^{-1/\xi}, & \xi \neq 0 \\ 1 - \exp\left(-\frac{x}{\sigma}\right), & \xi = 0 \end{cases} \tag{11}$$

och har stöd på $\left\{x > 0 \;\middle|\; 1 + \xi\frac{x}{\sigma} > 0\right\}$ [13, s. 75-76]. Se figur 3. Notera att fallet med $\xi = 0$ uppstår som ett gränsvärde när $\xi \longrightarrow 0$ av fördelningsfunktionen, precis som för GEV. GP kan även parametriseras med en ytterligare parameter, μ, kallad läge, som förskjuter hela fördelningen, men den är inte relevant för projektets tillämpning.

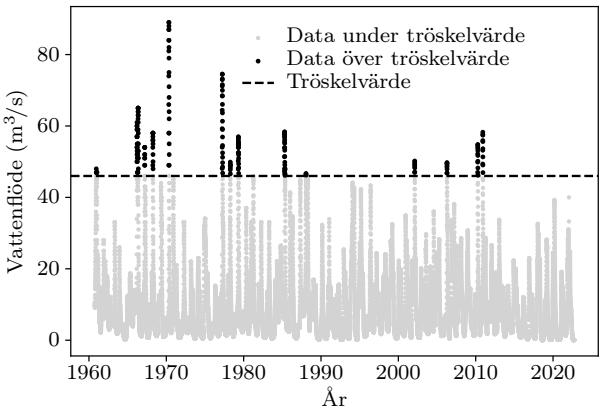


Figur 3: Sannolikhetstäthetsfunktion för GP-fördelningen med olika värden för ξ. Här är $\mu = 0$ och $\sigma = 1$.

2.1.4 Tröskelmetoden

Tröskelmetoden, även kallad *peaks over threshold*, analyserar endast mätvärden som överstiger en förbestämd tröskel, vanligtvis kallad u. Se figur 4. Om mätvärden är oberoende och likfördelade beskriver en tidshomogen Poissonprocess när överstigningar sker för stora tröskelvärden u [13, s. 128-130]. Tröskeln väljs till ett ”stort” tal eftersom asymptotiken med $u \longrightarrow \infty$ vill utnyttjas för samband 10 och resultatet om att överstigningar bildar en Poissonprocess.

Sannolikheten för överstigningar av specifika storlekar kan uppskattas genom att skatta parametrarna av GP-fördelningen och frekvensen λ från Poissonprocessen. GP-fördelningens parametrar kan skattas med maximum likelihood-skattningar. Parametern λ kan skattas med $\hat{\lambda} = \frac{N}{T}$ där T är den totala mättiden. Skattningen $\hat{\lambda}$ är väntevärdesriktig och dess varians går mot 0 då $T \longrightarrow \infty$. Se appendix A.2 för bevis.



Figur 4: Exempel för val av datapunkter i tröskelmetoden ifrån vattendraget Getebro.

Valet av tröskelvärdet innebär en avvägning mellan varians och systematiska fel i modellen. Ett för

lågt tröskelvärde riskerar att bryta den asymptotiska grunden för extremvärdesanalys och medföra systematiska fel i resultatet. Om ett för högt tröskelvärde väljs blir antalet överstigningar litet och då får skattningarna en stor varians [13]. Det finns två vanliga grafiska metoder för att göra avvägningen för tröskelvärdet. Den första grafiska metoden bygger på att beräkna medelvärdet för överstigningarna för olika tröskelvärden. Detta medelvärde kallas för *Mean Residual Life* på engelska, ofta förkortat MRL. Om MRL är linjärt i u efter ett visst tröskelvärde u_0 är överstigningarna av tröskelvärdena $u > u_0$ GP-fördelade [13, s. 78]. Ett lämpligt tröskelvärde kan alltså bestämmas genom att undersöka för vilket u_0 som MRL börjar vara linjärt i $u > u_0$. I praktiken minskar antalet datapunkter för större tröskelvärden och till sist blir variansen så stor att linjäriteten upphör.

Den andra vanliga grafiska metoden är att skatta GP-fördelningens parametrar, σ och ξ, för olika tröskelvärden u. Om värdena över tröskelvärdet följer en GP-fördelning bör skattningen av ξ vara konstant i u och σ linjär i u [13, s. 78]. Ett tröskelvärde väljs sedan i ett område där dessa förutsättningar uppfylls.

Grafiska metoder kräver en subjektiv bedömning för varje serie observationer som analyseras. I fall då många serier behandlas blir dessa processer tidskrävande. En metod som inte kräver ett beslut för varje enskild serie är att bestämma ett tröskelvärde som inkluderar en viss percentil av observationerna. Exempelvis är det vanligt att inkludera de 1 % till 10 % högsta observationerna. Den metoden har fördelen att ingen subjektiv bedömning behöver göras för varje dataserie.

2.2 Modeller med trend

Modellerna som hittills introducerats antar att alla observationer är oberoende och likfördelade. För att analysera trender i extremvärden antas det att parametrar i GEV-fördelningen och GP-fördelningen har ett tidsberoende. En GEV-modell med trend kan då uttryckas enligt

$$X_t \sim \text{GEV}(\mu_0 + \mu_1 t,\ e^{\phi_0+\phi_1 t},\ \xi_0 + \xi_1 t), \tag{12}$$

för $t \in \{0, \ldots, n-1\}$ där t representerar det t:te året. Alla X_t antas även vara oberoende. Fördelningarna parametriseras med $\phi = \ln(\sigma)$ istället för σ eftersom ett negativt σ-värde inte hade varit fysikaliskt. Parametrarna μ_0, $\sigma_0 = e^{\phi_0}$ och ξ_0 representerar parametrarna för GEV-fördelningen år $t = 0$. Se avsnitt 2.1.1. Trendparametrarna μ_1, ϕ_1 och ξ_1 representerar hur mycket dessa parametrar förändras per år.

GP-modellen med trend kan beskrivas med

$$X_t \sim \text{GP}(e^{\phi_0+\phi_1 t},\ \xi_0 + \xi_1 t), \tag{13}$$

för $t \in \{0, ..., n-1\}$ där t fortfarande representerar det t:te året. Parametrarna ϕ_0 och ξ_0 representerar parametrarna för GP-fördelningen år $t = 0$ i enlighet med avsnitt 2.1.3. Parametrarna ϕ_1 och ξ_1 representerar förändringen för respektive parameter per år.

2.3 Trend i frekvensen λ i Poissonprocessen

En tidshomogen Poissonprocess är ett starkt antagande för tröskelmetoden och kan ifrågasättas om man misstänker att det finns trender i frekvensen av överstiganden. Det är möjligt att överstigningar blir mer eller mindre vanliga över tid. För att modellera detta används en icke-homogen Poissonprocess. Observationer innan tid T är då Poissonfördelade med parameter $\Lambda(T)$ för

$$\Lambda(T) = \int_0^T \lambda(t)\,\mathrm{d}t. \tag{14}$$

Processens frekvens $\lambda(t)$ antas förändras över tid. Det naturliga valet är

$$\lambda(t) = e^{\lambda_1 t+\lambda_0} \tag{15}$$

eftersom λ nödvändigtvis är positiv, så en linjär trend skulle inte vara fysikalisk i alla tider. Parametern λ reparametriseras alltså likt hur σ blir ϕ i avsnitt 2.2 när trender i parametern undersöks. Om inget annat sägs antas Poissonprocessen vara tidshomogen och därför inte ha någon trend. Parametern λ refererar endast till frekvensen i en tidshomogen Poissonprocess. För en icke-homogen Poissonprocess används istället $\lambda(t)$, λ_0 och λ_1.

2.4 Maximum likelihood-skattning

Maximum likelihood-skattning, förkortat ML-skattning, är en parameterskattningsmetod för slumpvariabler $X_1, X_2, \ldots, X_n$. Metoden används när själva slumpvariablerna är kända, men parametrarna som beskriver variablernas fördelningsfunktion är okända. När ML används i praktiken är $X_1, X_2, \ldots, X_n$ ofta experimentiella datapunkter som man vill tillskriva en fördelningsfunktion.

Varje variabel X_i antas ha en känd fördelning givet en parametervektor θ. Sannolikhetstäthetsfunktionen för X_i är därför $f_i(X_i;\theta)$. Ofta är X_i likfördelade, så att $f_1 = f_2 = \cdots = f_n$. Två exempel är om alla X_i är normalfördelade, så ger det $\theta = (\mu, \sigma^2)$ eller om de är GEV-fördelade ger det $\theta = (\mu, \sigma, \xi)$. Slumpvariablerna behöver dock inte ha samma fördelningsfunktion, bara de är beroende av samma θ.

ML-skattningen skattar vilken parametervektor θ som med störst sannolikhet ger upphov till de givna variablerna $X_1, X_2, \ldots, X_n$. Sannolikheten att observera specifika $X_1, X_2, \ldots, X_n$ som en funktion av parametervektorn kallas ”likelihood”. Den gemensamma täthetsfunktionen är produkten av täthetsfunktionerna $f_i(X_i;\theta)$ om alla X_i är oberoende. Om $h(X_1, \ldots, X_n;\theta)$ är den gemensamma täthetsfunktionen är likelihood-funktionen

$$L(\theta) = h(X_1, \ldots, X_n;\theta) \overset{\text{ober}}{=} \prod_{i=1}^{n} f_i(X_i;\theta) \tag{16}$$

[14, s. 317].

Ett θ som maximerar $L(\theta)$ för givna $X_1, \ldots, X_n$ benämns maximum likelihood-skattning [14, s. 317].

Eftersom ln är en strikt växande funktion kommer maximum för

$$\ell(\theta) = \ln L(\theta) \overset{\text{ober}}{=} \sum_{i=1}^{n} \ln f_i(X_i;\theta) \tag{17}$$

vara samma θ som för maximeringen av $L(\theta)$ [14, s. 317]. Det är vanligtvis lättare att hitta maximum för en summa än för en produkt och därför används ofta $\ell(\theta)$ för att skatta det sanna θ [14, s. 317]. Maximum likelihood-skattningen av θ betecknas $\hat{\theta}$.

Maximum likelihood-skattningen har flera eftertraktade statistiska egenskaper. Ett första exempel är funktionell invarians. Förutsatt att g är en injektiv funktion och $\hat{\theta}$ är maximum likelihood-skattningen av θ är också $g(\hat{\theta})$ en ML-skattning av $g(\theta)$ [14, s. 320].

Under rimliga antaganden är $\hat{\theta}$ också en konsistent skattning av θ [14, s 321]. En skattning $\hat{\theta}$ är konsistent om följande gäller:

$$\forall \varepsilon > 0 \;\; \mathrm{P}\left(|\hat{\theta} - \theta| > \varepsilon\right) \xrightarrow[n \longrightarrow \infty]{} 0 \tag{18}$$

där n är antalet observationer och θ betecknar den sanna parametern [14, s 297].

Asymptotiskt är ML-skattningen, under det vanliga antagandet att Fisherinformationen existerar, skattningen med högst effektivitet, även kallad *efficiency* [14, s. 321]. En skattnings effektivitet är ett mått på hur stor varians skattningsvariablerna har, där lägre varians innebär en högre effektivitet.

ML-skattningen $\hat{\theta}$ är under samma antaganden också asymptotiskt väntevärdesriktig. Detta innebär att väntevärdet av $\hat{\theta}$ är den sanna parametern då antalet observationer $n \longrightarrow \infty$ [14, s 321]. Dessutom är ML-skattningar under vanliga förhållanden asymptotiskt normalfördelade [18, s. 83].

Existensen av en ML-skattning för GEV-fördelningen är bevisad i en begränsad, sluten omgivning till den sanna parametervektorn θ för $\xi > -1$. Under samma förutsättning med en begränsad, sluten omgivning är θ asymptotiskt normalfördelad, förutsatt att $\xi > -\frac{1}{2}$ [19, s. 2]. ML-skattning kan också tillämpas på GP-fördelningen [20].

2.5 Hypoteser och tester

I projektet används flera olika modeller som ger olika skattningar. Nollhypotestning används för att bestämma vilka resultat som är statistiskt signifikanta och vilka modeller som är relevanta. I varje test antas en nollhypotes som försöker förkastas genom ett test. Om testet ger ett resultat som är tillräckligt överraskande, alltså mindre sannolikt än den etablerade signifikansnivån, och ger bevis för vår alternativa hypotes förkastas nollhypotesen. För test av nollskilda parametrar är nollhypotesen att parametrarna faktiskt är 0.

2.5.1 Test av nollskilda parametrar

Antag att vi har en GEV-modell eller GP-modell med trend som i avsnitt 2.2. Det är intressant att undersöka om trendparametrarna är statistiskt signifikant skilda från 0. För GEV-modellen är det även intressant att testa om formparametern ξ är skild från 0. Nedan demonstreras det hur ett sådant test utförs. Antag att vi har en ML-skattning $\hat{\theta}$ med en standardavvikelse $\hat{\theta}_{SE}$ för en parameter θ. På grund av den asymptotiska normaliteten av ML-skattningar är

$$\hat{\theta} \sim \mathcal{N}(\theta,\ \hat{\theta}_{SE}^2) \tag{19}$$

och under nollhypotesen $H_0 : \theta = 0$ är därför

$$\frac{\hat{\theta}}{\hat{\theta}_{SE}} \sim \mathcal{N}(0, 1). \tag{20}$$

Ett p-värde för ett tvåsidigt test kan därmed bestämmas genom att anta att H_0 är sann och se att

$$p = \mathrm{P}\left(|z| < |X|\right) = 1 - \mathrm{P}\left(-|z| < |X| < |z|\right) = 1 - \left(\Phi(|z|) - \Phi(-|z|)\right) \tag{21}$$

$$= 1 - \left(\Phi(|z|) - (1 - \Phi(|z|))\right) = 2(1 - \Phi(|z|)), \tag{22}$$

där $z = \frac{\hat{\theta}}{\hat{\theta}_{SE}}$, $X \sim \mathcal{N}(0, 1)$ och Φ är standardnormalfördelningens fördelningsfunktion. Sedan kan p-värdet användas för ett test. Testet förkastar nollhypotesen om $p < \alpha$ på signifikansnivå α.

2.5.2 Anderson-Darlingtest

Anderson-Darlingtestet är ett godhetstest som används för att uppskatta hur väl datan passar en specifik fördelning. Testet bygger på att uppskatta skillnaden mellan den skattade fördelningen $F(x)$ och den empiriska kumulativa fördelningen $F_n(x)$ [21]. För en given fördelningsfunktion $F(x)$ är den empiriska kumulativa fördelningsfunktionen

$$F_n(x) = \frac{|\{x_i | x_i \leq x\}|}{n}. \tag{23}$$

Det betyder att F_n beskriver fördelningsfunktionen för en slumpvariabel som är fördelad enligt observationernas frekvenser. Testet ger ett p-värde för att testa nollhypotesen. Nollhypotesen är att datan kommer från en specifik fördelning, vanligtvis en skattad fördelning, och alternativhypotesen

är att datan inte kommer från den specifika fördelningen. Om $p < \alpha$ förkastas nollhypotesen på signifikansnivå α. För en djupare förklaring av proceduren för att räkna fram p-värden hänvisas intresserade läsare till appendix A.3.

2.5.3 Benjamini-Hochbergproceduren

Benjamini-Hochbergproceduren är en procedur som omvandlar p-värden för att ta hänsyn till de andra testerna. Antag att vi har m nollhypoteser $H_1, \ldots, H_m$ som ska testas och säg att tester av dessa gav p-värden $p_1, \ldots, p_m$. Detta försäkrar oss bara att testerna individuellt ger ett falskt positivt resultat med en signifikansnivå α. Det här är missvisande när man har många tester. Antag att nollhypoteserna är oberoende och att vi har $m > 1/\alpha$. I så fall förväntar vi oss åtminstone ett positivt resultat, även om alla nollhypoteser är sanna. För fler hypotestester ökar antalet förväntade falska positiva resultat.

Proceduren ser till att väntevärdet av andelen falska positiva resultat av alla positiva resultat är mindre än α [22, s. 3]. Väntevärdet för denna andel kallas FDR och står för *false discovery rate* [23, s. 1]. Det antas att hypoteserna är oberoende eller är positivt korrelerade [22, s. 4].

Följande är en beskrivning av proceduren. Först sorteras p-värdena i stigande ordning $p_{(1)} \leq \cdots \leq p_{(m)}$ med deras motsvarande nollhypoteser betecknade $H_{(1)}, \ldots, H_{(m)}$. Då blir de uppdaterade p-värdena

$$p_{(k)}^{\mathrm{BH}} = \min_{i:k \leq i} \left(p_{(i)} \frac{m}{i} \right) \tag{24}$$

[23, s. 5]. Nollhypotesen $H_{(k)}$ kan därmed förkastas om $p_{(k)}^{\mathrm{BH}} < \alpha$ med försäkringen att FDR för alla hypoteserna är mindre än α. Det innebär att om hela undersökningen, alltså observationer, datainsamling, skattning, tester och till slut Benjamini-Hochberg, hade gjorts om många gånger hade medelvärdet av andelen falska positiva resultat från alla studier varit mindre än α för tillräckligt många studier.

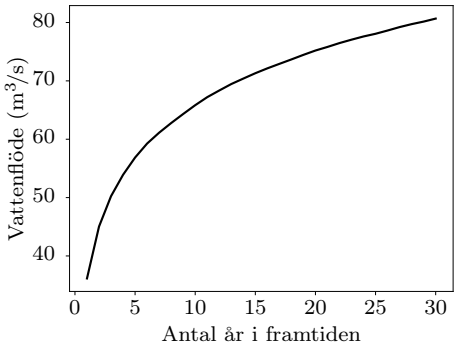
2.6 Återkomstnivåplott

SMHI:s definition av återkomstperiod, även kallad återkomstid eller *return period*, är följande [9].

> Återkomsttid är ett mått på hur ofta en ovanlig händelse kan förväntas. Med återkomsttid menas att händelsen i genomsnitt inträffar eller överträffas en gång under denna tid.

Återkomstnivå, även kallad *return level*, är storleken av den största händelse som förväntas inträffa under en viss återkomstperiod [13, s. 81]. Återkomstperioden är lika med $1/p$ år, givet att p är sannolikheten att återkomstnivån överskrids ett visst år [13, s. 49]. Notera att detta är baserat på ett antagande om stationäritet eftersom p inte är tidsberoende.

En återkomstnivåplott, även kallad *return level plot*, plottar återkomstperioder mot återkomstnivåer. Återkomstperioder plottas på x-axeln. Se figur 5 för vattendraget Getebro som ett exempel. Ur figuren går det exempelvis att avläsa att ett vattenflöde på $70\,\mathrm{m}^3/\mathrm{s}$ förväntas ske en gång under 15 år.



Figur 5: Återkomstnivåplott för vattendraget Getebro, enligt block maxima-modellen.

Återkomstnivåer har slutna analytiska uttryck. Återkomstnivån för t år, $\mathfrak{N}_t$, är

$$\mathfrak{N}_t = \mathrm{E}\left[\max\{X_1, \ldots, X_t\}\right], \tag{25}$$

där $X_1, \ldots, X_t \sim \mathrm{GEV}(\mu, \sigma, \xi)$. Eftersom GEV är maxstabil, se ekvation 5, är $\max\{X_1, \ldots, X_t\} \sim \mathrm{GEV}(\mu^*, \sigma^*, \xi^*)$. Notera att

$$\mathrm{E}\left[\max\{X_1, \ldots, X_t\}\right] = \begin{cases} \mu^* + \sigma^*\gamma, \ \xi^* = 0 \text{ [24, s. 11-12]} \\ \infty, \ \xi^* \geq 1 \text{ [25, s. 58]} \\ \mu^* + \frac{\sigma^*}{\xi^*}(\Gamma(1-\xi^*) - 1), \ \text{annars [26]} \end{cases} \tag{26}$$

där γ är Euler-Mascheronikonstanten och Γ är gammafunktionen.

2.7 Alternativ till återkomstnivåplott för modeller med trend

Eftersom återkomstnivåplottar inte fungerar för modeller med trend har vi själva utvecklat nya plottar. Vi kallar dem projektionsnivåplott respektive prediktionsnivåplott.

2.7.1 Projektionsnivåplott

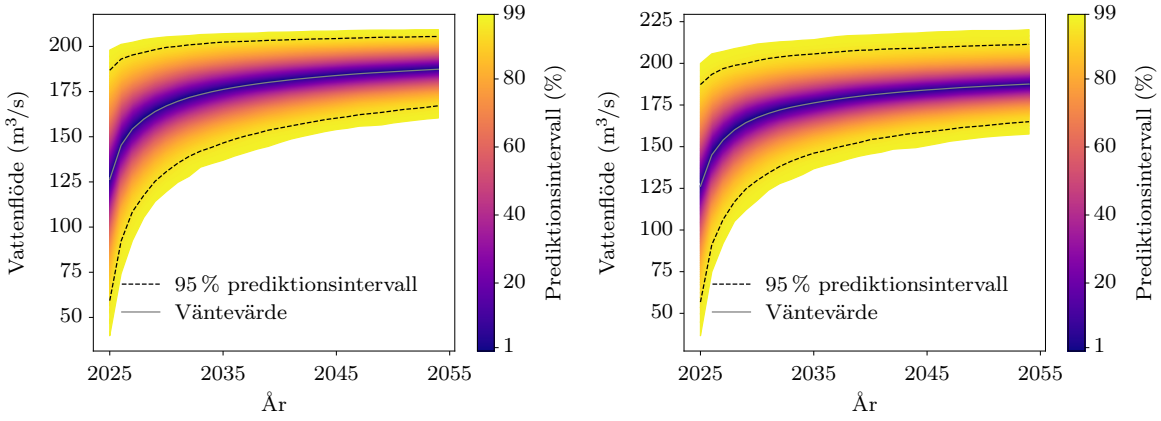
En återkomstperiod är inte väldefinierad för en modell med trend eftersom sannolikheten att en händelse händer om t år ändras från år till år. En projektionsnivåplott utgår därför från ett specifikt startår och listar år på x-axeln och nivåer på y-axeln. Till skillnad från en återkomstnivåplott visar den specifika år på x-axeln och inte år framåt. Till exempel skulle det stå ”2026” istället för ”2”. En projektionsnivåplott kan innehålla den förväntade framtida maximala nivån observerat fram till varje år, likt en återkomstnivåplott, men kan även innehålla prediktionsintervall. Med tillräckligt många olika prediktionsintervall bildas en gradient som tillåter läsaren att bedöma hur den maximala överstigningen fram till varje år är fördelad. Se figur 6. Notera att prediktionsintervallen inte tar hänsyn till modellosäkerhet eller parameterosäkerhet, utan endast stokasticiteten i modellen.

2.7.2 Prediktionsnivåplott

En prediktionsnivåplott är en projektionsnivåplott som har inkorporerat parameterosäkerhet. Detta kan göras Bayesianskt genom att ha en a-priorifördelning för parametrarna till modellen. Här presenterar vi en psuedo-Bayesiansk metod som bör vara funktionellt ekvivalent för stora datamängder.

Antag en GEV-modell med trend som i avsnitt 2.2. Låt nu $\theta = (\mu_0, \phi_0, \xi_0, \mu_1, \phi_1, \xi_1)$ vara en parametervektor. Givet en prior som är tillräckligt glatt, positiv runt den sanna parametervektorn och ett antal andra tekniska krav kommer både a-posteriorifördelningen och en maximum likelihood-skattning av θ att bli asymptotiskt normalfördelade och likfördelade när antalet observationer går mot oändligheten [27, s. 184-185]. A-posteriorifördelningen blir då asymptotiskt $\mathcal{N}^6(\hat{\theta}, \hat{\Sigma})$, där $\hat{\theta}$

är parametervektorn med ML-skattningarna och $\hat{\Sigma}$ är inversen till Fisherinformationen för ML-skattningarna. Därför kan den asymptotiska a-posteriorifördelningen skattas med hjälp av endast en maximum likelihood-skattning med mycket data. Prediktionsnivåplottens prediktionsintervall och väntevärde reflekterar istället den prediktiva a-posteriorin, till skillnad från projektionsplotten som grundar sig i ML-skattningarna och prediktionsintervall som uppstår från dem. Se figur 6 för en jämförelse mellan projektionsnivå- och prediktionsnivåplott. Ur prediktionsnivåplotten kan man exempelvis se att modellen förutser en 99 % sannolikhet att det maximala flödet observerat mellan 2025 och 2035 är mellan cirka 135 och 210 kubikmeter per sekund.



(a) Projektionsnivåplott genererad för Ankarvattnetmed block maxima-modellen utan trend.

(b) Prediktionsnivåplott genererad för Ankarvattnetmed block maxima-modellen utan trend

Figur 6: Projektions- och prediktionsnivåplottar för Ankarvattnet. Parameterosäkerheten leder till större prediktionsintervall i prediktionsnivåplotten.

3 Metod

Projektet använde framför allt block maxima- och tröskelmetoden för analys av vattenflöden. Datan som använts i rapporten är inhämtad från SMHI och har behandlats i Python för att sedan ha analyserats i R. Graferna som presenteras är skapade med Python. Signifikansnivån som använts för alla tester var 5 %.

3.1 Modeller

I avsnitt 2.2 introduceras block maxima och tröskelmetoden med trend för att analysera vattendrag. Projektet analyserade trender i μ och ϕ, men inte i ξ eftersom det ansågs mindre intressant. Parametrarna uttrycks som $\mu = \mu_0 + \mu_1 t$ och $\phi = \phi_0 + \phi_1 t$, där t är tid i år. För modeller utan trend skattades endast μ_0 och ϕ_0, medan både μ_1 och ϕ_1 sattes till noll. För modeller med trender skattades μ_0 och μ_1 eller ϕ_0 och ϕ_1 beroende på vilken modell som valdes. Det är intressant att veta om det finns trender i lägesparametrarna eller skalparametrarna och därför används fem olika modeller: block maxima utan trender, block maxima med trender i μ, block maxima med trender i ϕ, tröskelmetoden utan trender och tröskelmetoden med trender i ϕ. Modellerna har sammanställts i tabell 1. För alla modeller valdes år $t = 0$ till 1960 eftersom det är där mätdatan börjar.

Tabell 1: Tabell med de olika modellerna och hur de skiljer sig åt. Nej innebär att inga trender undersökts för parametern. Ja innebär att trender undersökts. N/A innebär att modellen saknar trendparametern.

Modellnamn	Trend i μ_1	Trend i ϕ_1	Trend i ξ_1	Fördelning
Block maxima utan trender	Nej	Nej	Nej	GEV
Block maxima med trender i μ	Ja	Nej	Nej	GEV
Block maxima med trender i ϕ	Nej	Ja	Nej	GEV
Tröskelmetod utan trender	N/A	Nej	Nej	GP
Tröskelmetod med trender i ϕ	N/A	Ja	Nej	GP

3.2 Datainsamling och databehandling

SMHI utför mätningar av vattenföring på över 200 platser i Sverige [28] och har därför valts som källa till data i rapporten. Datan från SMHI som används har ett flödesvärde per dygn och vattendrag angivet i kubikmeter per sekund. Under de första åren av den undersökta perioden var dessa värden baserade på manuell mätning och SMHI benämner dessa värden lågupplösta eftersom dygnsvärdet är baserat på en linjär interpolation mellan mätningarna. För det mesta skedde mätningarna dagligen, men under vintertid förekom det att mätningarna endast skedde veckovis. Under den undersökta perioden har insamlingsmetoden bytts ut mot realtidsmätare där dygnsvärdet är ett medelvärde över hela dygnet [4].

Datan är allmänt tillgänglig och har hämtats genom SMHI:s API, *Application Programming Interface*, för meterologiska observationer [29] med hjälp av programmeringsspråket Python. Datan har sorterats enligt specifika krav för att kunna fastställa eventuella trender, inklusive ett krav på långsiktig insamling. Endast mätstationer som varit aktiva utan avbrott ifrån 1960 och framåt använts. För att analyserna för de olika vattendragen skall vara jämförbara betraktas endast data mellan 1:a oktober 1960 till 30:e september 2022.

3.3 Dataanalys

Programmeringsspråket R användes för att göra de statistiska analyserna. R valdes eftersom det är byggt för statistisk databehandling och innehåller många statistikpaket. I analysen användes paketen Ismev, NHPoisson, gnFit, extRemes och lubridate. Den insamlade datan användes sedan för att skatta parametrarna och deras trender med en maximum-likelihood skattning, beskriven i avsnitt 2.4.

Block maxima-metoden användes för att generera en fördelning som approximeras till en GEV-fördelning. Baserat på information som presenteras i avsnitt 1.1 och teorin i avsnitt 2.1.2 har datamängden delats upp i vattenår. Vattenår minskar risken för att höga flöden i slutet av året ska bidra till stora flöden även kommande år. Vidare ger längden år en intuitiv förståelse för förekomsten av extrema händelser. Årsindelningens påverkan kommer inte att undersökas vidare enligt de specificerade avgränsningarna.

Valet av tröskelvärde är avgörande för hur väl GP-fördelningen representerar datamängden i tröskelmetoden. Enligt avsnitt 2.1.4 är det vanligt att inkludera 1 % till 10 % högsta observationerna. Tröskelvärdet valdes så att 2 % av datan kommer att överstiga tröskelvärdet. Detta ger i genomsnitt 7,3 överskridande värden per år.

En utmaning med tröskelmetoden är att överstigningar kan hamna i kluster, definierat som en följd mätpunkter som alla är överstigningar omgivet av mätpunkter som inte är överstigningar. Det kan vara större sannolikhet att det kommer en överstigning direkt efter en annan överstigning. För att ta hänsyn till detta används så kallad ”deklustring”. Deklustring bygger på att identifiera kluster och sedan välja ut endast det största överstigandet i klustret och förkasta resten. Sedan utförs tröskelmetodsanalys på de kvarvarande överstigningarna. Detta har en teoretisk grund och är en vanlig procedur [13, s. 99-100, 177].

Trend i en icke-homogen Poissonprocess kan ML-skattas med paketet NHPoisson förutsatt att modellen från avsnitt 2.3 stämmer. Den deklustrade datan användes för att skatta $\lambda(t)$. NHPoisson använder dagar som tidsenhet dagar som tidsenhet, så frekvensen mäts i dagar tillskillnad från oss som använder år. Detta justeras därför i efterhand. Den dagliga trenden multiplicerades med 365.25 för att ge λ_1 och $\ln(365.25)$ adderades till λ_0 som ges av NHPoisson. Detta justerar enheten till per år från per dag. För bevis av detta, se appendix A.4.

Modellerna utan trender testades med Anderson-Darlingtestet från avsnitt 2.5.2. Paketet gnFit utförde testet för att bedöma deras lämplighet. Teorin från avsnitt 2.5.1 användes för att testa om trendparametrarna och parametern ξ är nollskilda. Benjamini-Hochbergproceduren från avsnitt 2.5.3 användes för att säkerställa att FDR är maximalt 5 % så länge som hypoteserna inte är positivt beroende. Proceduren tillämpades på p-värden från Anderson-Darlingtesterna, men separat för varje modell. Testerna för nollskilda parametrar Benjamini-Hochbergjusteras också, men det gjordes separat för varje modell och parameter.

3.4 Monte Carlo-sampling

Under projektet utvecklades en Monte Carlo-sampler i R för extremvärdesprocesser med trender från avsnitt 2.2. Monte Carlo-samplingen bygger på simulering av ett stort antal framtida års maximala nivåer. Sedan skattas väntade maximala årsnivåer, sannolikheter för händelser och prediktionsintervall genom medelvärden, andelar och kvantiler av den simulerade datan. Denna sampler möjliggör skattningen av extrema händelser givet modellparametrar, antingen enskilda eller tillsammans med kovarianser mellan dem. Om samplern blir given en kovariansmatris, se avsnitt 2.7.2, används den för att uppskatta en asymptotisk a-posteriorifördelning för modellparametrarna, vilket sedan används för sampling. Om samplern inte blir given en kovariansmatris används endast ML-skattningarna som modellparametrar. Samplern kan endast hantera skattningar av λ för den homogena Poissonprocessen i tröskelmetodmodellerna och utför ingen sampling av den. Samplern stöder inte trender i frekvensparametern λ. Samplern användes för att uppskatta sannolikheten för framtida vädervarningar för alla modeller samt för att generera projektionsnivåplottar och prediktionsnivåplottar från avsnitt 2.7. Den använder alltid parametersampling från a-posteriorifördelningen för alla vädervarningsrisker. Alla parametrar i modellen samplas, även om de inte bedömts statistiskt signifikant skilda från noll i tester eftersom samplern verkar Bayesianskt. Tröskelmetodmodellerna i samplern använder λ före deklustring för att korrekt bedöma sannolikheten av överstigningar. Den mest generella modellen med trender i flera parametrar på samma gång från avsnitt 2.2 stöds, men i projektet används endast en trendparameter i taget.

Projektionsnivåplotten, till skillnad från återkomstnivåplotten, kan inte utnyttja maxstabiliteten av GEV-fördelningen eftersom parametertrenden innebär att varje observation inte är likfördelad. Vårt program använder istället Monte Carlo-samplern för att kunna hantera parametertrender.

Prediktionsplotten använder ett stickprov av parametervektorns a-posteriorifördelningen istället för ML-skattningarna av parametrarna för varje simulerad framtid.

3.5 Riskförändring

Riskförändringsmåttet har konstruerats genom att välja det största vattenflödet de senaste 5, 25 respektive 50 åren som bas för en gul, orange eller röd vädervarning. Därefter samplas vattenflödesmaximum för det kommande året utifrån modellen med Monte Carlo-samplern från avsnitt 3.4, vilket ger en sannolikhet för att en gul, orange eller röd vädervarningsnivå kommer att överstigas inom det kommande året. Många stickprov av vattenflödesmaximum från modellen 30 år i framtiden tas sedan, vilket uppskattar en sannolikhet för att en gul, orange eller röd vädervarningsnivå överstigs under ett år, 30 år i framtiden. Kvoten mellan dessa två sannolikheter minus ett benämns riskförändring. Exempelvis betyder en riskförändring av 100 % att risken dubblats, en riskförändring av 0 % ingen riskförändring och en riskförändring av -50 % att risken halveras, alla 30 år i framtiden jämfört med nu.

Modellernas parametrar är svårtolkade. Ta exempelvis vattendraget Getebro och modellen block maxima med trender i ϕ. Där är $\phi_1 = -0{,}0150$. Detta värdets koppling till trender i översvämningsrisker är inte uppenbart. För att tydliggöra vad de anpassade parametrarna innebär för översvämningsrisker presenteras därför riskförändringar. Riskförändringen indikerar hur risken av en händelse förändrats efter en tidsperiod. Händelserna i fråga är vattenflöden som idag motsvarar gul, orange och röd vädervarning. Det nuvarande årets risk jämförs med risken 30 år i framtiden. Att vattendraget Getebro har en riskförändring av -50 %, -62 % och -70 % för gul, orange respektive röd vädervarning innebär alltså att vädervarningarnas sannolikheter om 30 år enligt modellen kommer att minska med 50 %, 62 % respektive 70 % jämfört med dagens sannolikhet. Risken för en gul vädervarning beskriver sannolikheten för en gul vädervarning eller värre. På samma sätt är risken för orange vädervarning orange eller värre. Detta sätt att räkna valdes för att undvika missförstånd och göra informationen lättare att tolka. En negativ riskförändring i orange hade annars kunnat indikera att risker minskar eller exempelvis att nästan alla extrema vattenflöden blir extremt stora så de klassificeras som röda. Dessa två fall beskriver två väldigt olika processer.

4 Resultat

Nedan belyses signifikanta och sammanfattande resultat för att besvara projektets syfte och frågeställningar. Intresserar läsaren sig för vattendragsspecifika resultat eller mer detaljrika resultat uppmuntras hen läsa i bilagan till kandidatarbetet [30]. Där presenteras alla resultat för alla modeller och vattendrag som framkommit under arbetet.

4.1 Modellers giltighet

Benjamini-Hochbergjusterade Anderson-Darlingtester för nollhypotesen att den skattade fördelningen faktiskt genererade mätdatan presenteras i tabell 2. I endast fyra vattendrag förkastades nollhypotesen med en signifikansnivå på 5 % för block maxima-modellen. För tröskelmetoden förkastades 26 utav 54 nollhypoteser. Detta antyder att tröskelmetoden inte är lika lämplig som block maxima-metoden för dessa datamängder och därför läggs större vikt på presentation av resultat från block maxima-modellerna.

Tabell 2: Anderson-Darlingtest för modellernas giltighet. Benjamini-Hochbergproceduren har utförts separat för varje typ av test i varje modell. För varje modell testades 54 vattendrag.

Anderson-Darlingtest med Benjamini-Hochberg	Antal förkastade modeller
Block maxima-modellen utan trender	4
Tröskelmetoden utan trender	26

Notera att modellerna utan trender är specialfall av modellerna med trend, med trendparametrar satta till 0. Därför är en välpassning av block maxima utan trender bevis för välpassandet av block maxima med trender i exempelvis ϕ.

4.2 Trender i modeller

Tabell 3 visar resultat från test av nollhypotesen att trendparametrarna är noll med en signifikansnivå på 5 %. Ingen signifikant trend i lägesparametern kunde observeras i något vattendrag för block maxima-modellen, förutsatt att modellen block maxima med trend i μ från tabell 1 är sann. Däremot förkastades nollhypotesen att ϕ saknade trend, alltså att $\phi_1 = 0$, i 10 vattendrag för samma modell. Det innebär att det finns 10 vattendrag med statistisk signifikant trend i ϕ givet att modellen block maxima med trend i ϕ är korrekt. För tröskelmetoden med trend i ϕ kunde nollhypotesen att $\phi_1 = 0$ förkastas 36 gånger, vilket indikerar statistiskt signifikanta trender i 36 vattendrag, givet att modellen är sann.

Tabell 3: Testresultat för nollhypotesen att en trendparameter är 0. Benjamini-Hochbergproceduren har utförts separat för varje typ av test i varje modell på de 54 vattendragen.

Nollhypotes	Antal förkastningar
$\mu_1 = 0$ i block maxima med trend i μ	0
$\phi_1 = 0$ i block maxima med trend i ϕ	10
$\phi_1 = 0$ i tröskelmetoden med trend i ϕ	36

Av de 54 vattendrag som testades var det endast fyra som visade signifikanta trender i frekvensen λ. De vattendrag som visade signifikanta trender redovisas i appendix B.

4.3 Vädervarningar

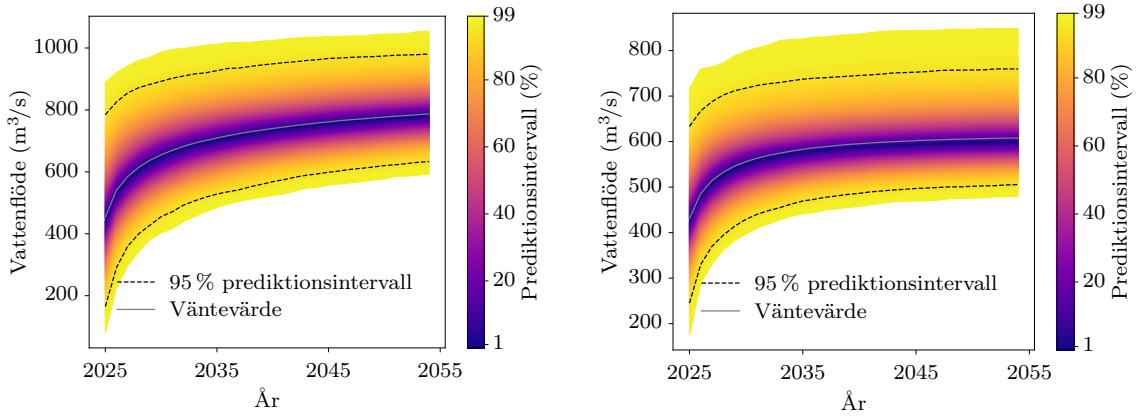
Tabell 4 visar trenden i ϕ och vädervarningarnas förändringar för block maxima-modellen med trend i ϕ. Notera att de flesta riskförändringarna är negativa vilket motsvarar att de mest extrema händelserna blir alltmer sällsynta. Övre Lansjärv och Motala KRV visar däremot en positiv trend.

Mer detaljerade uppgifter för parameterskattnigar och riskförändringar för samtliga vattendrag återfinns i appendix B. Rööån har ett saknat värde i tabellen. Detta uppstår från att den röda risknivån modelleras som omöjlig idag och om 30 år, vilket innebär att en division med noll utförs i uträkningen.

Tabell 4: Block maxima med trend i ϕ

	Parametervärden		Riskförändring		
Vattendrag	ϕ_0	ϕ_1	Gul varning	Orange varning	Röd varning
Getebro	2,9877	−0,0152	−50 %	−62 %	−70 %
Karats	3,9753	−0,0160	−67 %	−68 %	−100 %
Lannavaara	5,3082	−0,0119	−70 %	−82 %	−77 %
Motala krv	2,9871	0,0003	0 %	4 %	30 %
Niavve	4,9263	−0,0135	−61 %	−100 %	−100 %
Rööån	1,8504	−0,0193	−92 %	−90 %	-
Storuman 2	5,0297	−0,0242	−66 %	−88 %	−83 %
Sundstorp	3,0171	−0,0121	−20 %	−69 %	0 %
Övre hyndevad	3,3014	−0,0037	−96 %	−97 %	−100 %
Övre lansjärv	3,6927	0,0058	177 %	232 %	173 %

Figur 7 visar prediktionsnivåplotten för vattendraget Lannavaara med respektive utan trend i ϕ. Utan trender i ϕ, se figur 7a, observeras en ökad varians i det predikterade vattenflödet jämfört med modellen med trender i figur 7b. Notera skillnaden i den vertikala axeln. Det är också märkvärt att det maximala värdet inte har en betydlig ökning med tiden. Detta beror på att GEV-fördelningen i vattendraget är av typ $\xi < 0$ och ger vattenflödena ett absolut maxvärde. Minns från avsnitt 2.1.1 att denna typ har en övre gräns på stödet av sannolikhetstäthetsfunktionen.



(a) Prediktionsnivåplott genererad för Lannavaara med block maxima-modellen utan trend.

(b) Prediktionsnivåplott genererad för Lannavaara med block maxima-modellen med trend i ϕ.

Figur 7: Prediktionsnivåplottar för Lannavaara. Trenden i ϕ är negativ, vilket resulterar i lägre vattenflöden i prediktionsplotten. Notera att prediktionsnivåerna är något smalare i modellen med trend i ϕ.

5 Diskussion

Godhetstesterna förkastade block maxima-modellen 4 av 54 gånger medan tröskelmetoden förkastades 26 gånger. Därför anses tröskelmetoden passa dåligt och fokus läggs istället på resultaten från block maxima-modellerna. Trender i skalparametern, vilket motsvarar spridningen i storleken av extrema vattenflöden, beskrivs exponentiellt med parametern ϕ. Det fanns statistiskt signifikanta trender i parametern ϕ för tio vattendrag i block maxima-modellen, alltså 19 % av vattendragen, där åtta trender var negativa och två var positiva. Inga statistiskt signifikanta trender i lägesparametern μ upptäcktes i block maxima-modellerna. I endast 4 av de 54 vattendragen fanns det bevis för trender i frekvensen λ av extrema vattenflöden.

Tidigare resultat ifrån SMHI upptäckte inga vidsträckta trender i svenska vattendrags översvämningar. Detta stämmer överens med vår rapports resultat. SMHI har jämfört genomsnitt av de högsta flödena över tid istället för att använda extremvärdesanalys som utnyttjas i denna rapport. De har även använt andra urvalskriterier som att vattendragen ska vara relativt oreglerade, men de har inte tagit lika stor hänsyn till saknad mätdata [11].

Hodkins m.fl. [31] undersökte europeiska och nordamerikanska floder och fann inga generella trender i extrema vattenflöden. Trenderna undersöktes genom att analysera antalet överstiganden över anpassade tröskelvärden över tid. En senare studie analyserade samma floder genom att undersöka trender i parameteranpassningen för GEV- och GP-fördelningen, men fann då signifikanta trender i en minoritet av vattendragen. Trots parametertrenderna fanns ingen tydlig trend i de modellerade återkomstnivåerna [32]. Datamängden som användes för de båda undersökningarna hade andra krav på vattendragen än denna studie, bland annat låg mänsklig inverkan och urbaniseringsgrad, vilket resulterade i att endast nio svenska vattendrag undersöktes [31].

Ett anmärkningsvärt resultat i vår rapport är att tröskelmetoden utan trender passade dåligt. En möjlig förklaring är att stora kluster i datan innebar att deklustringen tog bort för mycket data från analysen. För få datapunkter leder till stor varians i anpassningen av GP-modellen. Exempelvis är mätdatan från vattendragen Möckeln, Gimdalsby samt Övre Lansjärv starkt klustrade vilket innebär att det endast fanns ett lågt antal mätpunkter kvar att skatta med efter deklustring. I analyser av tornados, vars processer tenderar att bilda stora kluster, används även den negativa binomialfördelningen som modell [33, s. 1]. Användningen av denna fördelning som ett alternativ för svenska vattendrag föreslås undersökas i framtida forskning. Få datapunkter är också en möjlig förklaring till att vissa skattade parametrar för tröskelmetodmodellerna har standardavvikelser som är flera tiopotenser mindre än parameterns värde. En annan orsak till den dåliga anpassningen av tröskelmetoden skulle kunna vara att tröskelvärdet valdes genom en på förhand bestämd percentil, vilket är en kritiserad metod som ändå används i praktiken [34, s. 41]. Den asymptotiska regeln som föreslås av Loretan och Philips [35] ger en tröskelandel på 98,5 % för vår mängd data, mycket likt det som valdes. Ytterligare en möjlig förklaring till att tröskelmetoden utan trender inte passade är att tröskeln var konstant för varje vattendrag. En tidsvarierande tröskel är ett alternativ [13, s. 136]. Att ha ett konstant tröskelvärde i data med trender kan medföra att bias och varians förändras över tid och resultera i att modellen inte passar.

Block maxima-modellen påverkas inte av de här förhållandena. Eftersom block maxima-modellen endast studerar årsmaxima kommer klustring i datan inte påverka modellen meningsfullt och till skillnad från svårigheterna i val av tröskelvärde så är valet av vattenår väletablerat inom hydrologi.

Ett av projektets avgränsningar var att inte undersöka skillnader i vattendragens reglering, exempelvis om det finns en damm eller ett vattenkraftverk i vattendraget som kan störa det naturliga flödet. Eftersom modellerna var dåligt anpassade till datan så undersöktes möjliga felkällor. Det visade sig att modellanpassningarna förskastas i högre grad för reglerade vattendrag, vilket demonstreras i tabell B.3. Därför bör det tas hänsyn till om vattendragen är reglerade om projektet upprepas. Exempelvis är Övre Hyndevad reglerat och har små standaravvikelser i block maxima-modellen med trender i ϕ, se tabell B.1. En möjlig förklaring till detta är att det finns en damm precis vid mätstationen [36].

En faktor som kan påverka riktigheten av resultatet är SMHI:s insamlingsmetod. Insamlingsmeto-

den ändrades under den undersökta perioden, vilket kan innebära att datan från början av perioden inte är lämplig att jämföra med datan från slutet. Under 1900-talet gick SMHI över till automatiska mätningar från att ha mätt för hand [28].

De specifika värdena för trendanpassningarna och deras innebörd är svårttolkade utan att diskutera vädervarningsrisker. De negativa riskförändringarna tyder på att extrema vattenflöden blir mindre vanliga, vilket skulle innebära att risken för översvämmad mark i närhet till vattendragen samt skador på infrastruktur minskar.

I flera vattendrag minskar risken för röda varningsnivåer med 100 %. En anledning till detta är att anpassningen gav en negativ formparameter ξ, vilket leder till en omvänd Weibullfördelning och en sådan modell innebär att det existerar ett absolut maxflöde. I och med de negativa trenderna i parametern ϕ minskar detta maxflöde varje år enligt modellen och leder till sist till att den röda varningsnivån ligger över modellens maxflöde. Det kan diskuteras hur verklighetstroget ett maximalt vattenflöde är, men en effekt av den omvända Weibullfördelningen som inte är rimlig är att den möjliggör godtyckligt stora negativa vattenflöden. Eftersom riskförändringen skattats 30 år i framtiden har parametertrenderna fått stora effekter. En kortare riskförändringstid, exempelvis 10 år, kanske hade representerat verkliga förhållanden bättre.

Vädervarningsmodellerna bygger på Monte Carlo-samplern som asymptotiskt har en korrekt implementering av parameterosäkerhet. Detta är dock bara en approximation eftersom ingen apriorifördelning utgicks från och mängden data är ändlig. Samplern kan heller inte hantera modellosäkerhet. Läsaren bör därför vara skeptisk gällande risksannolikheterna, speciellt de väldigt sällsynta röda och dagliga riskerna.

Skattade sannolikheter för extremt sällsynta händelser är väldigt känsliga för modellantaganden och antalet samples i Monte Carlo-samplern. Därmed bör mindre vikt läggas på dessa resultat. Variansen av samplerns resultat analyseras inte, men minst en miljon samples tas i varje riskberäkning. Minst tio miljoner samples tas i riskberäkningar för tröskelmetodmodellerna.

I projektet användes Benjamini-Hochbergproceduren för varje modell och test för att korrekt multipeltesta vattendrag. Detta tar inte multipeltestningen mellan modeller och tester i åtanke. Så länge läsare är medvetna om innebörden av detta och inte misstolkar signifikansen av resultatet anser författarna att detta är acceptabelt. Ett alternativt sätt att presentera resultaten hade varit att tillämpa Benjamini-Hochbergproceduren på alla p-värden i hela projektet tillsammans och hade då garanterat FDR mindre än α bland alla tester tillsammans. Detta hade tyvärr också varit ointuitivt och gjort olika modeller och testers förkastning beroende av varandra. Nollhypoteserna för alla tester mellan alla vattendrag antogs inte vara positivt korrelerade eftersom Benjamini-Hochberg användes. Ett sätt att undvika detta antagande hade varit att istället använda Benjamini-Yekutieli, en modifiering av Benjamini-Hochberg [22, s. 1169]. Användandet av Benjamini-Hochberg för Anderson-Darlingtestresultat kan kritiseras då Anderson-Darling förkastar om modellerna inte passar. Benjamini-Hocbergproceduren kontrollerar därför för falska negativa resultat och inte falska positiva resultat för vår hypotes.

Det är möjligt att en övergripande trend finns, men att denna inte kunnat påvisats. En tänkbar orsak är att projektet inte hade nog med styrka för att upptäcka trenden. Styrka är sannolikheten att nollhypotesen förkastas givet en specifik alternativhypotes. Trenden kan vara så liten att testet given datan inte är starkt nog.

5.1 Samhälleliga och etiska aspekter

Projektets resultat kan användas för att minimera skador på miljön och ekosystemet. Viktiga ekosystemtjänster från vattendrag inkluderar biologisk mångfald, dricksvatten, vattenflödesreglering och rekreation. Översvämningars effekt på biologisk mångfald beror på vilka områden som drabbas. Jordbruksmark innehåller näringsämnen som kväve och fosfor som riskerar att spridas vid översvämningar. Avrinning från industri- och stadsmiljöer ökar läckaget av närningsämnen och föroreningar. Detta belastar inte bara ekosystemtjänsteren utan även den biologiska mångfalden [37].

Projektet har även möjliga samhälleliga konsekvenser för invånare i översvämningskänsliga områden. Resultatet ökar förståelsen för översvämningsrisker och kan leda till utveckling av skyddsstrategier och beslut om stadsplanering. Samhälleliga effekter delas vanligtvis in i direkta och indirekta effekter samt materiella och immateriella. Direkt materiella effekter inkluderar fysiska skador på infrastruktur och byggnader samt materiella förluster. Immateriella direkta effekter omfattar förlust av liv och skador. Dessutom finns indirekta effekter såsom störningar i leveranskedjan och trafik, samt långsiktiga hälsokonsekvenser [38]. Genom att identifiera trender i förekomsten av extrema vattenflöden i Sverige kan projektet bidra till att minska konsekvenser på både miljön och samhället.

Endast offentlig data från SMHI har använts, vilket inte riskerar att konfidentiell data läcks. Insamling av data har skett utan att SMHI meddelats, men med strävan att minimera anrop till deras server för att minska kostnader och undvika överbelastning av deras hemsida.

Resultatet har publicerats offentligt i Chalmers kandidatarbetesdatabas [39] utan anonymisering eftersom den inte innehåller personuppgifter eller känslig information. Datan är även publicerad på GitHub. SMHI anser att deras databaser är en samhällsresurs och strävar efter att göra all användbar data tillgänglig för allmänheten i digital form [40].

Extremvärdesanalysen kan inte hantera icke-kvantifierbara risker, alltså trendbrytande extrema händelser med andra orsaker än de hittills observerade. Ett exempel av ett icke-kvantifierbart extremt vattenflöde skulle vara om en damm brast. De observerade extrema vattenflödena i Sverige har helt andra orsaker och skulle alltså inte informera oss om översvämningsrisken från dammbristningen.

För att undvika missförstånd och felaktiga tolkningar som kan leda till överdrivningar eller förminskanden av uppfattningen av problematiken, har all teori och metod noggrant presenterats. Vidare har resultaten granskats och diskuterats för att säkerställa tydlighet i slutsatserna samt identifiera felkällor.

5.2 Slutsatser

Det verkar inte finnas vidsträckta trender i extrema vattenflöden i Sverige. Tio vattendrag har trender i skala, varav åtta är negativa trender. I flera vattendrag modellerades risken för SMHI:s röda vädervarningar helt försvinna om 30 år, vilket inte är ett lika tillförlitligt resultat som avsaknaden av generella trender.

Ett förslag till vidare forskning hade varit att använda modellen på data från 1960-2000 och se hur väl den predikterade datan från 2001-2022. Datan skulle alltså delas upp i träningsdata och valideringsdata. Ytterligare föreslår vi att vidare undersöka tröskelmetoden för svenska vattendrag genom att använda andra metoder för tröskelval. Dessutom är det intressant att studera hur reglering av vattendrag påverkar de extrema flödena samt tillämpbarheten av extremvärdesanalys i detta fall.

Referenser

[1] SMHI. *Mänsklig påverkan på sjöar och vattendrag*. URL: https://www.smhi.se/kunskapsbanken/hydrologi/mansklig-paverkan/mansklig-paverkan-pa-sjoar-och-vattendrag-1.178251. (Läst 2024-01-25).

[2] David Francis, Henry Hengeveld m. fl. *Extreme weather and climate change*. Environment Canada Ontario, 1998.

[3] SMHI. *Vattenföring*. URL: https://www.smhi.se/kunskapsbanken/hydrologi/vattenforing/vattenforing-1.6705. (Läst 2024-04-15).

[4] SMHI. *SMHIs vattenföringsmätningar*. URL: https://www.smhi.se/kunskapsbanken/hydrologi/matning-av-flode-och-vattenstand/smhis-vattenforingsmatningar-1.80833. (Läst 2024-04-15).

[5] USGS. *What is Hydrology?* URL: https://www.usgs.gov/special-topics/water-science-school/science/what-hydrology#Hydrology. (Läst 2024-03-04).

[6] Don Johnstone och William Perry Cross. "Elements of Applied Hydrology". I: The Ronald Press Company, 1949, s. 102–103.

[7] USGS. *Explanations for the National Water Conditions*. URL: https://water.usgs.gov/nwc/explain_data.html. (Läst 2024-03-04).

[8] SMHI. *SMHIs vädervarningar*. URL: https://www.smhi.se/kunskapsbanken/meteorologi/varningar-och-meddelanden/smhis-vadervarningar-1.167835. (Läst 2024-04-09).

[9] SMHI. *Återkomsttider*. URL: https://www.smhi.se/kunskapsbanken/klimat/extremer/aterkomsttider-1.89085. (Läst 2024-04-09).

[10] Axel Bronstert. "Floods and Climate Change: Interactions and Impacts". I: *Risk Analysis* 23.3 (2003), s. 545–557. DOI: https://doi.org/10.1111/1539-6924.00335. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/1539-6924.00335.

[11] SMHI. *Klimatindikator - vattenflöde*. URL: https://www.smhi.se/klimat/klimatet-da-och-nu/klimatindikatorer/klimatindikator-vattenflode-1.202365. (Läst 2024-01-25).

[12] Christopher P. Konrad. *Returns to Investment in Higher Education*. United States Geological Survey, 2003. DOI: https://doi.org/10.3133/fs07603.

[13] Stuart Coles. *An Introduction to Statistical Modeling and Extreme Values*. Springer, 2001. ISBN: 1852334592.

[14] Peter Olofsson och Mikael Andersson. *Probability, Statistics and Stochastic Processes*. Second. John Wiley & Sons, Inc., 2012. ISBN: 9780470889749.

[15] Dipak K Dey och Jun Yan. *Extreme value modeling and risk analysis: methods and applications*. CRC Press, 2016.

[16] A. A. Balkema och L. de Haan. "Residual Life Time at Great Age". I: *The Annals of Probability* 2.5 (1974), s. 792–804. DOI: 10.1214/aop/1176996548. URL: https://doi.org/10.1214/aop/1176996548.

[17] James Pickands III. "Statistical Inference Using Extreme Order Statistics". I: *The Annals of Statistics* 3.1 (1975), s. 119–131. DOI: 10.1214/aos/1176343003. URL: https://doi.org/10.1214/aos/1176343003.

[18] Wang D Abdelzaher T Kaplan L. *Social Sensing Building Reliable Systems on Unreliable Data*. Morgan Kaufmann, 2015. ISBN: 978-0-12-800867-6.

[19] L. Zhang och B. A Shaby. *ASYMPTOTIC POSTERIOR NORMALITY OF THE GENERALIZED EXTREME VALUE DISTRIBUTION*. 2023.

[20] Scott D Grimshaw. "Computing Maximum Likelihood Estimates for the Generalized Pareto Distribution". I: *Technometrics* 35.2 (1993), s. 185.

[21] T. W. Anderson och D. A. Darling. "A Test of Goodness of Fit". I: *Journal of the American Statistical Association* 49.268 (1954), s. 765–769. ISSN: 01621459. URL: http://www.jstor.org/stable/2281537 (hämtad 2024-04-10).

[22] Yoav Benjamini och Daniel Yekutieli. "The Control of the False Discovery Rate in Multiple Testing under Dependency". I: *The Annals of Statistics* (2001).

[23] Daniel Yekutieli och Yoav Benjamini. "Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics". I: *Journal of Statistical Planning and Inference* (2001).

[24] Norman L Johnson, Samuel Kotz och Narayanaswamy Balakrishnan. *Continuous univariate distributions, volume 2*. Vol. 289. John wiley & sons, 1995.

[25] Jan Beirlant m. fl. *Statistics of extremes: theory and applications*. John Wiley & Sons, 2006.

[26] G Muraleedharan, C Guedes Soares och Claudia Lucas. "Characteristic and moment generating functions of generalised extreme value distribution (GEV)". I: *Sea level rise, coastal engineering, shorelines and tides* (2011), s. 269–276.

[27] C. C. Heyde och I. M. Johnstone. "On Asymptotic Posterior Normality for Stochastic Processes". I: *Journal of the Royal Statistical Society. Series B (Methodological)* 41.2 (1979), s. 184–189. ISSN: 00359246. URL: http://www.jstor.org/stable/2985031 (hämtad 2024-05-03).

[28] SMHI. *SMHIs vattenföringsmätningar*. URL: https://www.smhi.se/kunskapsbanken/hydrologi/matning-av-flode-och-vattenstand/smhis-vattenforingsmatningar-1.80833. (Läst 2024-03-04).

[29] SMHI. *SMHI Open Data API Docs - Meterological Observations*. URL: https://opendata.smhi.se/apidocs/metobs/index.html. (Läst 2024-01-30).

[30] E. Ahlström m. fl. *Bilaga till kandidatarbete*. https://github.com/VincentHarbander/Vattenflode. 2024.

[31] Glenn A. Hodgkins m. fl. "Climate-driven variability in the occurrence of major floods across North America and Europe". I: *Journal of Hydrology* 552 (2017), s. 704–717. ISSN: 0022-1694. DOI: https://doi.org/10.1016/j.jhydrol.2017.07.027. URL: https://www.sciencedirect.com/science/article/pii/S002216941730478X.

[32] Axel Ström. "Trends in Flooding in Europe and North America-an Extreme Value Approach". I: *LUNFMS-3081-2018* (2018).

[33] HCS Thom. "Tornado probabilities". I: *Monthly weather review* 91.10 (1963), s. 730–736.

[34] Carl Scarrott och Anna MacDonald. "A review of extreme value threshold estimation and uncertainty quantification". I: *REVSTAT-Statistical journal* 10.1 (2012), s. 33–60.

[35] Mico Loretan och Peter CB Phillips. "Testing the covariance stationarity of heavy-tailed time series: An overview of the theory with applications to several financial datasets". I: *Journal of empirical finance* 1.2 (1994), s. 211–248.

[36] Hjälmarens vattenförbund. *Hyndevadsdammen*. URL: https://hjalmarensvattenforbund.se/hyndevadsdammen/. (Läst 2024-05-02).

[37] L. Bergström m. fl. *Klimatförändringar och biologisk mångfald – Slutsatser från IPCC och IPBES i ett svenskt perspektiv*. KLIMATOLOGI Nr 56. SMHI och Naturvårdsverket, 2020. URL: https://www.smhi.se/polopoly_fs/1.164056!/Klimatologi_56%20Klimatf%C3%B6r%C3%A4ndringar%20och%20biologisk%20m%C3%A5ngfald.pdf.

[38] Philip Bubeck, Antje Otto och Juergen Weichselgartner. "Societal Impacts of Flood Hazards". I: *Oxford Research Encyclopedia of Natural Hazard Science* (2017). URL: https://doi.org/10.1093/acrefore/9780199389407.013.281.

[39] Chalmers tekniska högskola AB. *Examensarbeten för kandidatexamen*. URL: https://odr.chalmers.se/collections/43957f85-4338-4f3f-a665-e44d1c58629a. (Läst 2024-05-02).

[40] SMHI. *SMHIs datapolicy*. https://www.smhi.se/omsmhi/policys/datapolicy. Läst: 18:02 6/2-2024. 2023.

[41] Gemai Chen och N. Balakrishnan. "A General Purpose Approximate Goodness-of-Fit Test". I: *Journal of Quality Technology* 27.2 (1995), s. 154–161. DOI: 10.1080/00224065.1995.11979578. eprint: https://doi.org/10.1080/00224065.1995.11979578. URL: https://doi.org/10.1080/00224065.1995.11979578.

[42] Ralph B. D'Agostino. *GOODNESS-OF-FIT TECHNIQUES*. Springer, 2001. ISBN: 1852334592.

A Appendix 1 – Teori

A.1 Bevis av fördelning av maximum av oberoende GEV-fördelningar

Låt $X_1, \ldots, X_n \sim \text{GEV}(\mu, \sigma, \xi)$ och F vara fördelningsfunktionen för dessa. Vi bevisar nedan att $\max\{X_1, \ldots, X_n\} \sim \text{GEV}(\mu^*, \sigma^*, \xi^*)$ för

$$\begin{aligned} \mu^* &= \mu + \sigma \ln n, \\ \sigma^* &= \sigma, \\ \xi^* &= \xi \end{aligned} \tag{27}$$

om $\xi = 0$. Annars är

$$\begin{aligned} \mu^* &= \mu + \frac{\sigma}{\xi}(n^\xi - 1), \\ \sigma^* &= n^\xi \sigma, \\ \xi^* &= \xi. \end{aligned} \tag{28}$$

Observera först att

$$\begin{aligned} \text{P}\left(\max\{X_1, \ldots, X_n\} \leq x\right) &= \text{P}\left(X_1 \leq x, \ldots, X_n \leq x\right) \\ &\overset{\text{ober}}{=} \text{P}\left(X_1 \leq x\right) \ldots \text{P}\left(X_n \leq x\right) = (F(x))^n. \end{aligned} \tag{29}$$

Nu falluppdelar vi. Antag först att $\xi = 0$. Då är

$$\begin{aligned} (F(x))^n &= \left(\exp\left(-e^{-\frac{x-\mu}{\sigma}}\right)\right)^n = \exp\left(-ne^{-\frac{x-\mu}{\sigma}}\right) = \exp\left(-e^{\ln n}e^{-\frac{x-\mu}{\sigma}}\right) \\ &= \exp\left(-e^{\ln n - \frac{x-\mu}{\sigma}}\right) = \exp\left(-e^{-\frac{x-\mu-\sigma\ln n}{\sigma}}\right) = \exp\left(-e^{-\frac{x-(\mu+\sigma\ln n)}{\sigma}}\right), \end{aligned} \tag{30}$$

vilket kan identifieras som fördelningsfunktionen för en GEV-fördelning med parametrar

$$\begin{aligned} \mu^* &= \mu + \sigma \ln n, \\ \sigma^* &= \sigma, \\ \xi^* &= \xi. \end{aligned} \tag{31}$$

Antag istället att $\xi \neq 0$. Vi kan därför se att

$$\begin{aligned} (F(x))^n &= \left(\exp\left(-\left(1 + \xi\frac{x-\mu}{\sigma}\right)^{-\frac{1}{\xi}}\right)\right)^n = \exp\left(-n\left(1 + \xi\frac{x-\mu}{\sigma}\right)^{-\frac{1}{\xi}}\right) \\ &= \exp\left(-\left(n^{-\xi} + \frac{n^{-\xi}\xi}{\sigma}(x-\mu)\right)^{-\frac{1}{\xi}}\right) = \exp\left(-\left(1 + (n^{-\xi} - 1) + \frac{\xi}{n^\xi\sigma}(x-\mu)\right)^{-\frac{1}{\xi}}\right) \\ &= \exp\left(-\left(1 + \frac{\xi}{n^\xi\sigma}\left(x - \mu + \frac{n^\xi\sigma}{\xi}(n^{-\xi} - 1)\right)\right)^{-\frac{1}{\xi}}\right) \\ &= \exp\left(-\left(1 + \frac{\xi}{n^\xi\sigma}\left(x - \mu + \frac{\sigma}{\xi}(1 - n^\xi)\right)\right)^{-\frac{1}{\xi}}\right) \\ &= \exp\left(-\left(1 + \frac{\xi}{n^\xi\sigma}\left(x - \mu - \frac{\sigma}{\xi}(n^\xi - 1)\right)\right)^{-\frac{1}{\xi}}\right) \\ &= \exp\left(-\left(1 + \frac{\xi}{n^\xi\sigma}\left(x - \left(\mu + \frac{\sigma}{\xi}(n^\xi - 1)\right)\right)\right)^{-\frac{1}{\xi}}\right), \end{aligned} \tag{32}$$

vilket kan identifieras som fördelningsfunktionen för en GEV-fördelning med parametrar

$$\begin{aligned}\mu^* &= \mu + \frac{\sigma}{\xi}(n^\xi - 1),\\ \sigma^* &= n^\xi \sigma,\\ \xi^* &= \xi.\end{aligned} \tag{33}$$

A.2 Bevis för skattningsegenskaper av $\hat{\lambda}$

Låt N vara antalet observationer av en Poissonprocess under mättid T. Vi bevisar nu att

$$\hat{\lambda} = \frac{N}{T} \tag{34}$$

är väntevärdesriktig och att när mättiden T går mot oändligheten kommer variansen av skattningen gå mot 0.

Observera att

$$N \sim \text{Poi}(\lambda T). \tag{35}$$

Låt $\hat{\lambda} = N/T$ och notera att

$$\text{E}\left[\hat{\lambda} - \lambda\right] = \text{E}\left[\frac{N}{T}\right] - \lambda = \frac{1}{T}\text{E}\left[N\right] - \lambda = \frac{1}{T}\lambda T - \lambda = 0. \tag{36}$$

Därför ser vi att $\hat{\lambda}$ är väntevärdesriktig. Notera att

$$\text{Var}\left(\hat{\lambda}\right) = \text{Var}\left(\frac{N}{T}\right) = \frac{1}{T^2}\text{Var}\left(N\right) = \frac{1}{T^2}\lambda T = \frac{\lambda}{T} \xrightarrow[T \longrightarrow \infty]{} 0. \tag{37}$$

Variansens gränsvärde är därmed visat.

A.3 Uträkning av p-värden för Anderson-Darlingtest

Proceduren kommer från en artikel skriven av Chen and Balakrishnan [41]. Först sätts de skattade parametrarna $\hat{\theta}$ in i $v_i = F(x_i; \hat{\theta})$ där $F(x; \theta)$ är fördelningen som testas och x_i är datan sorterad i ökande storleksordning. Sedan formuleras $y_i = \Phi(v_i)$ för att räkna ut $u_i = \Phi(\frac{y_i - \bar{y}}{\sigma_y})$ som slutligen ger

$$W_n^2 = -n - n^{-1}\sum_{i=1}^{n}(21 - \ln(u_i) + (2n + 1 - 2i)\ln(1 - u_i)). \tag{38}$$

Tabellen A.1 [42, s. 127] använder sedan värdet som fås från

$$W^* = W_n^2(1 + \frac{0.75}{n} + \frac{2.25}{n^2}) \tag{39}$$

för att räkna ut p-värdet.

Tabell A.1: Tabell som visar hur p-värden beräknas.

Intervall	p-värde
$W^* < 0{,}2000$	$-1 - \exp\left(-13{,}4360 + 101{,}1400W^* - 223{,}7300\left(W^*\right)^2\right)$
$0{,}2000 \leq W^* < 0{,}3400$	$1 - \exp\left(-8{,}3180 + 42{,}7960W^* - 59{,}9380\left(W^*\right)^2\right)$
$0{,}3400 \leq W^* < 0{,}6000$	$\exp\left(0{,}9177 - 4{,}2790W^* - 1{,}3800\left(W^*\right)^2\right)$
$0{,}6000 \leq W^*$	$\exp\left(1{,}2937 - 5{,}7090W^* + 0{,}0186\left(W^*\right)^2\right)$

A.4 Bevis av enhetsbyte för trend i λ

Minns att $\alpha = 365,25$ är konversionsfaktorn från år till dagar. Låt oss beteckna den dagliga frekvensen som $\tilde{\lambda}$ och

$$\tilde{\lambda}(d) = e^{\tilde{\lambda}_1 d + \tilde{\lambda}_0} \tag{40}$$

för en godtyckligt dag d. Vi är istället intresserade av $\lambda(t)$, den årliga frekvensen, för ett godtyckligt år t. Eftersom det finns i genomsnitt α dagar per år är $d = \alpha t$ och $\lambda(t) = \alpha\tilde{\lambda}(d)$. Eftersom

$$\lambda(t) = e^{\lambda_1 t + \lambda_0} \tag{41}$$

blir

$$e^{\lambda_1 t + \lambda_0} = \lambda(t) = \alpha\tilde{\lambda}(d) = \alpha e^{\tilde{\lambda}_1 d + \tilde{\lambda}_0} = e^{\ln(\alpha)} e^{\alpha\tilde{\lambda}_1 t + \tilde{\lambda}_0} = e^{\alpha\tilde{\lambda}_1 t + \tilde{\lambda}_0 + \ln(\alpha)}. \tag{42}$$

Därmed kan λ_0 och λ_1 identiferas i termer av $\tilde{\lambda}_0$ och $\tilde{\lambda}_1$ som

$$\begin{aligned} \lambda_1 &= \alpha\tilde{\lambda}_1, \\ \lambda_0 &= \tilde{\lambda}_0 + \ln(\alpha). \end{aligned} \tag{43}$$

B Appendix 2 – Tabeller

B.1 Värden för λ

Tabell B.1: Trender i frekvensen, $\lambda(t) = e^{\lambda_1 t + \lambda_0}$, för signifikanta vattendrag.

Vattendrag	λ_0	λ_1
Landösjön nedre	3,833	−0,007
Munkedal 2	3,808	−0,005
Röån	1,850	−0,019
Övre Hyndevad	3,301	−0,004

B.2 Värden för ξ

Tabell B.2: Testresultat för nollhypotesen att $\xi = 0$. Benjamini-Hochbergproceduren har utförts separat för varje typ av test och modell. För varje modell testades 54 vattendrag.

Nollhypotes	Antal förkastningar
$\xi = 0$ i BM med trend i μ	29
$\xi = 0$ i BM med trend i ϕ	27
$\xi = 0$ i PoT med trend i ϕ	15

B.3 Reglerade vattendrag

I tabellen B.3 redovisas icke-Benjamini-Hochberjusterade Anderson-Darlingtester för vattendragen baserade på om de är reglerade eller inte. Av de 54 undersökta vattendragen var 31 oreglerade och 23 reglerade.

Tabell B.3: Anderson-Darlingtest för modellernas giltighet för alla vattendrag, oreglerade vattendrag och reglerade vattendrag.

Anderson-Darling test	Antal förkastade modeller
Block maxima-modellen	13/54
Tröskelmetoden	28/54
Block maxima-modellen oreglerat	5/31
Tröskelmetoden oreglerat	10/31
Block maxima-modellen reglerat	8/23
Tröskelmetoden reglerat	18/23

C Appendix 3 – Kod

C.1 LaTeX-generatorn för bilagan

```
# Vincent
# 2024

# pdflatex main.tex --shell-escape

import numpy as np, pandas as pd

all_data = "R/All␣data.csv"

df = pd.read_csv(all_data, sep=";", encoding="utf-8")

result = ""
example = ""

with open("input.txt", mode="r", encoding="utf-8") as file:
    example = file.read()

for i in range(len(df)):
    curr = df.iloc[i,:]

    raw = curr["name"][:-4]  # Remove the .csv ending
    split_list = raw.split("_")
    underscoredrealname = "_".join(split_list[:-1])
    name = "␣".join(split_list[:-1])
    station = split_list[-1]

    chunk = example.replace("underscored", raw)
    .replace("realname", name).replace("stationnr", station)

    # These are replaced first because "lambda0" is a
    substring of "lambda0_se", which is inconvenient!
    chunk = chunk.replace("␣lambda0_se", "␣\\num{" + str(curr["lambda0_se"]) + "}")
    chunk = chunk.replace("␣lambda1_se", "␣\\num{" + str(curr["lambda1_se"]) + "}")
    chunk = chunk.replace("␣lambda1_p", "␣\\num{" + str(curr["lambda1_p"]) + "}")
    chunk = chunk.replace("␣lambda1_bh", "␣\\num{" + str(curr["lambda1_bh"]) + "}")
```

```
    for param in curr.keys():
        if param == "post_datapts" or str(curr[param]) == "nan":
            chunk = chunk.replace("␣" + param, "␣" + str(curr[param]))
        elif param in ["yellow_level", "orange_level", "red_level"]:
            chunk = chunk.replace(param, "\\qty{" + str(curr[param])
            + "}{\\meter\\cubed\\per\\second}")
        else:
            chunk = chunk.replace("␣" + param, "␣\\num{" + str(curr[param]) + "}")

    result += chunk

with open("LaTeX.txt", mode="w", encoding="utf-8") as file:
    file.write(result)
```

C.2 Monte Carlo-samplern

```
# Vincent Harbander
# 2024

library(eva)
library(mvtnorm)

sampleParams <- function(paramMean, paramCov, number=1) {
  return(mvtnorm::rmvnorm(number, paramMean, paramCov))
}

sampleGEV <- function(year, zeroYear, loc0, phi0, shape0, loc1=0, phi1=0,
    shape1=0) {
  # This function returns a sample from the maximum flow of a specific year
      , simulated

  # year is the current year you want to sample from
  # check levelResultsGEV for the meaning of the rest of the arguments

  mu <- loc0+loc1*(year-zeroYear)
  sigma <- exp(phi0+phi1*(year-zeroYear))
  xi <- shape0+shape1*(year-zeroYear)

  # eva::rgevr can take lists of parameters, BUT returns dependent samples!
       A bug in the library! Always gives the same sample for the same
      parameters!
  return(eva::rgevr(1, 1, mu, sigma, xi))
}

sampleVectorGEV <- function(year, zeroYear, loc0, phi0, shape0, loc1=c(),
    phi1=c(), shape1=c()) {
  # This function returns a VECTOR OF INDEPENDENT of samples from the
      maximum flow of a specific year, simulated

  # NOTE: The parameter arguments here should be vectors, unlike in
      sampleGEV!

  # year is the current year you want to sample from
  # check levelResultsGEV for the meaning of the rest of the arguments

  # Assumes all of phi0 and shape0 have the same length as loc0
  n <- length(loc0)
  res <- rep(0, n)

  # If no trend was specified, have it be 0
  if(length(loc1) == 0) {
```

```
    loc1 <- rep(0, n)
  }
  if(length(phi1) == 0) {
    phi1 <- rep(0, n)
  }
  if(length(shape1) == 0) {
    shape1 <- rep(0, n)
  }

  res <- rep(0, n)

  for(i in 1:n) {
    res[i] <- sampleGEV(year, zeroYear, loc0[i], phi0[i], shape0[i], loc1[i
        ], phi1[i], shape1[i])
  }

  return(res)
}


dataGEV <- function(T, year, zeroYear, loc0, phi0, shape0, loc1=0, phi1=0,
    shape1=0, cov=matrix(0, 6, 6), samples=10000) {
  # This function returns results of sample futures over T years

  # see the rest of the arguments from levelDataGEV

  # Sample parameters for each sample
  # paramVectors is a samples x dim(paramMeans) matrix
  paramVectors <- sampleParams(c(loc0, phi0, shape0, loc1, phi1, shape1),
      cov, samples)
  l0 <- paramVectors[,1]
  p0 <- paramVectors[,2]
  s0 <- paramVectors[,3]
  l1 <- paramVectors[,4]
  p1 <- paramVectors[,5]
  s1 <- paramVectors[,6]

  res <- matrix(0, samples, T)

  # The first one is necessarily max, after that we have to pick the max so
      far
  # t=1 is handled here
  res[,1] <- sampleVectorGEV(year+1, zeroYear, l0, p0, s0, l1, p1, s1)

  # Has +t to NOT include the current year. NOTE that is starts on 2 since
      1 is handled above
  if(T > 1) {
    for(t in 2:T) {
      # Pass vectors of parameters
      res[,t] <- sampleVectorGEV(year+t, zeroYear, l0, p0, s0, l1, p1, s1)
    }
  }

  return(res)
}

# Need to not just take a sample, but take t *timed* samples and take the
    max of those
# This takes a new future sample and returns the biggest one seen so far
levelDataGEV <- function(T, year, zeroYear, loc0, phi0, shape0, loc1=0,
    phi1=0, shape1=0, cov=matrix(0, 6, 6), samples=10000) {
```

```
  # This returns the samples from max{X_1,...,X_T} over times t=1:T in a
      matrix

  # phi=ln(sigma)
  # T is the time period, the number of years.
  # see the rest of the arguments from levelResultsGEV

  # Sample parameters for each sample
  # paramVectors is a samples x dim(paramMeans) matrix
  paramVectors <- sampleParams(c(loc0, phi0, shape0, loc1, phi1, shape1),
      cov, samples)
  l0 <- paramVectors[,1]
  p0 <- paramVectors[,2]
  s0 <- paramVectors[,3]
  l1 <- paramVectors[,4]
  p1 <- paramVectors[,5]
  s1 <- paramVectors[,6]

  res <- matrix(0, samples, T)

  # The first one is necessarily max, after that we have to pick the max so
       far
  # t=1 is handled here
  res[,1] <- sampleVectorGEV(year+1, zeroYear, l0, p0, s0, l1, p1, s1)

  # Has +t to NOT include the current year. NOTE that is starts on 2 since
      1 is handled above
  if(T > 1) {
    for(t in 2:T) {
      # Pass vectors of parameters
      sample <- sampleVectorGEV(year+t, zeroYear, l0, p0, s0, l1, p1, s1)
      # If we found something even bigger, update the new max
      # Perform an element-wise max
      res[,t] <- pmax(res[,t-1], sample)
    }
  }

  return(res)
}

levelResultsGEV <- function(timePeriod, startYear, zeroYear, loc0, phi0,
    shape0, loc1=0, phi1=0, shape1=0, cov=matrix(0, 6, 6), samples=10000,
    confidence=c(0.95)) {
  # This function returns a vector of average simulated return levels for
      every year in a given time period and a [confidence]-prediction
      interval for the level, all simulated

  # timePeriod (T) is the time period you want to check the return level
      for
  # startYear is the year you want to check from (not including startYear)
  # zeroYear is the year where location=loc0, at zeroYear+1 location=loc0+
      loc1, etc
  # loc refers to mu, the location
  # phi = ln(sigma), where sigma is the scale
  # shape refers to xi
  # confidence gives the function the confidence for the inside of the
      prediction interval
  # confidence is given as a list of decreasing confidences
  # samples gives how many samples are sampled PER YEAR to determine mean
      and prediction interval, more is better
  # The cov is the covariance matrix of the parameters listed as mu0, phi0,
```

```
      xi0, mu1, phi1, xi1
  # The parameters are sampled normally around the parameter with the given
      standard deviation

  resData <- levelDataGEV(timePeriod, startYear, zeroYear, loc0, phi0,
      shape0, loc1, phi1, shape1, cov, samples)

  means <- colMeans(resData)

  alpha <- 1-confidence

  pred <- matrix(0, 2*length(confidence), timePeriod)

  for(t in 1:timePeriod) {
    pred[,t] <- quantile(resData[,t], c(alpha/2, rev(1-alpha/2)))
  }

  # deparse.level just makes sure it doesn't create any labels for the
      dataframe
  return(rbind(means, pred, deparse.level=0))
  #return(c(means, pred))
}

levelPlotGEV <- function(timePeriod, startYear, zeroYear, loc0, phi0,
    shape0, loc1=0, phi1=0, shape1=0, cov=matrix(0, 6, 6), samples=10000,
    confidence=(0.95)) {
  # This function simply calls levelResultsGEV and handles the plotting,
      returning nothing
  # See levelResultsGEV for the meaning of the arguments

  years <- matrix(rep((startYear+1):(startYear+timePeriod), 3), nrow=3,
      byrow=TRUE)
  result <- levelResultsGEV(timePeriod, startYear, zeroYear, loc0, phi0,
      shape0, loc1, phi1, shape1, cov, samples, confidence)

  n <- length(confidence)
  ramp <- colorRampPalette(c("orange", "red"), alpha=TRUE)(n)
  colors <- c(ramp, rev(ramp))

  conf <- as.character(100*confidence)

  for(i in 1:length(conf)) {
    conf[i] <- paste(conf[i], "%", sep="")
  }

  # Plots COLUMNS
  matplot(t(years), t(result), type="l", lty=1,
          col=c("blue", colors), xlab="Year",
          ylab="Max␣encountered␣level", main="Prediction␣level␣plot")
  legend("bottomright", legend=c("Mean", conf[1:n]),
         col=c("blue", ramp),
         lty=1)
}

probAboveGEV <- function(level, timePeriod, startYear, zeroYear, loc0, phi0
    , shape0, loc1=0, phi1=0, shape1=0, cov=matrix(0, 6, 6), samples
    =1000000) {
  # Returns the probability of exceeding level within the specified time
      period
  # startYear is NOT included in the time period analyzed
  # See levelResultsGEV for the rest of the arguments
```

```
  # resData contains samples of the biggest occurrance within the time
      period
  tmp <- levelDataGEV(timePeriod, startYear, zeroYear, loc0, phi0, shape0,
      loc1, phi1, shape1, cov, samples)
  resData <- tmp[,timePeriod] # Pick out the last time, it's necessarily
      the max

  # Now we count exceedances!
  exceedances <- 0

  for(elem in resData) {
    if(elem > level) {
      exceedances <- exceedances + 1
    }
  }

  return(exceedances/samples)
}

sampleGP <- function(year, zeroYear, phi0, shape0, phi1=0, shape1=0) {
  # This function returns a sample from the maximum flow of a specific year
      , simulated

  # year is the current year you want to sample from
  # check levelResultsGEV for the meaning of the rest of the arguments

  sigma <- exp(phi0+phi1*(year-zeroYear))
  xi <- shape0+shape1*(year-zeroYear)

  # eva::rgpd is also bugged with parameters as vectors!!! Need to write my
       own vectorization!
  return(eva::rgpd(1, 0, sigma, xi))
}

sampleVectorGP <- function(year, zeroYear, phi0, shape0, phi1=c(), shape1=c
    ()) {
  # This function returns a VECTOR of INDEPENDENT samples of GP given a
      vector of parameters and years, simulated

  # See sampleGP for the arguments

  # Assumes shape0 has the same length as phi0
  n <- length(phi0)
  res <- rep(0, n)

  # If no trend was specified, have it be 0
  if(length(phi1) == 0) {
    phi1 <- rep(0, n)
  }
  if(length(shape1) == 0) {
    shape1 <- rep(0, n)
  }

  res <- rep(0, n)

  for(i in 1:n) {
    res[i] <- sampleGP(year, zeroYear, phi0[i], shape0[i], phi1[i], shape1[
        i])
  }
```

```
  return(res)
}

probAboveGP <- function(level, u, year, zeroYear, lambda, phi0, shape0,
    phi1=0, shape1=0, cov=matrix(0, 4, 4), samples=10000000) {
  # This function returns the probability that an occurrance above level
      happens during year using PoT (Peaks over threshold)

  # year is the EXACT year you want to check
  # ASSUMES level > u!
  # u is the threshold for PoT
  # cov is the covariance matrix of the parameters listed as phi0, xi0,
      phi1, xi1

  excessLevel <- level - u
  samplePoi <- rpois(samples, lambda) # Sample the number of threshold
      excesses for each sample future

  # Sample parameters as vectors!
  sampleVectors <- sampleParams(c(phi0, shape0, phi1, shape1), cov, samples
      )

  p0 <- sampleVectors[,1]
  s0 <- sampleVectors[,2]
  p1 <- sampleVectors[,3]
  s1 <- sampleVectors[,4]

  exceedances <- 0

  for(i in 1:samples) {
    n <- samplePoi[i]
    if(n == 0) { next }
    for(j in 1:n) { # For each occurance of this sample year
      # Note! Not sampleVectorGP!
      res <- sampleGP(year, zeroYear, p0[i], s0[i], p1[i], s1[i])

      # Is this high enough to be relevant?
      if(res > excessLevel) {
        exceedances <- exceedances + 1
        # Need to break here! We only care if an exceedance occurs or not.
        break
      }
    }
  }

  return(exceedances/samples)
}

probAboveDayGP <- function(level, u, year, zeroYear, lambda, phi0, shape0,
    phi1=0, shape1=0, cov=matrix(0, 4, 4), samples=365*1000000) {
  # This function returns the probability that an occurrance above level
      happens during A DAY using PoT (Peaks over threshold)

  # year is the EXACT year you want to check
  # ASSUMES level > u!
  # u is the threshold for PoT
  # cov is the covariance matrix of the parameters listed as phi0, xi0,
      phi1, xi1

  # Use Poisson thinning and increased default samples to compensate
  return(probAboveGP(level, u, year, zeroYear, lambda/365, phi0, shape0,
```

```
      phi1, shape1, cov, samples))
}

# Ankarvattnet_1537 enligt BM
#mu0 <- 106.005731007067
#mu1 <- 0.286752012672177
#sigma <- 33.8274083582117
#xi <- -0.335594561819737

# Made up, purely for testing
#cov <- matrix(0, 6, 6)
#cov[1, 1] <- 1
#cov[4, 4] <- 0.01
#cov[1, 4] <- 0.01
#cov[4, 1] <- cov[1, 4]

# Ankarvattnet_1537 enligt PoT
#lambda <- 0.01
#gpphi <- 3.4
#gpxi <- -0.23

# Made up, purely for testing
#gpcov <- matrix(0, 4, 4)
#gpcov[3,3] <- 0.3

#levelPlotGEV(timePeriod=30, startYear=2024, zeroYear=1960, loc0=mu0, phi0=
    log(sigma), shape0=xi, loc1=mu1, cov=cov, confidence=(19:1)/20)
```

C.3 Dataanalys

```
# Simon
# 2024

library(ismev)
library(extRemes)
library(gnFit)
source("projectedLevel.R")
library(NHPoisson)
library(lubridate)

# BM0 - Block maxima without trends
BM_function <- function(k, waterflow_data_transformad) {
  # Max per year
  year_max <- apply.yearly(waterflow_data_transformad, max)

  # The data needs to be in this form
  x <- matrix(ncol = 1, nrow = length(year_max$Value))
  x[,1] <- seq(1,length(year_max$Value),1)

  # ML estimates
  fit.gev_start_value <- gev.fit(year_max$Value, ydat= x, show = FALSE,
      method="BFGS")
  fit.gev <- gev.fit(year_max$Value, ydat= x, siglink = exp, show = FALSE,
      method="BFGS", muinit = fit.gev_start_value$mle[1], siginit = log(fit
      .gev_start_value$mle[2]), shinit = 0.1)

  estimates <- fit.gev$mle
  SError <- fit.gev$se

  # Covaraince matrix
  mat <- (fit.gev$cov)
```

```
cov <- matrix(0,6,6)
cov[1,1] = mat[1,1]
cov[2,1] = cov[1,2] = mat[2,1]
cov[3,1] = cov[1,3] = mat[3,1]
cov[2,2] = mat[2,2]
cov[3,2] = cov[2,3] = mat[3,2]
cov[3,3] = mat[3,3]

# Tests
ad <- gnfit(year_max$Value, "gev", pr = c(estimates[1], exp(estimates[2])
    , estimates[3]))$Apval
xi_p <- 2*(1 - pnorm(abs(estimates[3]/SError[3]), 0, 1))
test <- ifelse(0 > estimates[3]- 1.96*SError[3] & 0 < estimates[3]+ 1.96*
    SError[3], 0, estimates[3])

# Weather predictions
yellow_now = probAboveGEV(max(tail(year_max$Value, n=5)), 1, 2023, 1960,
    estimates[1], (estimates[2]), estimates[3], cov=cov, samples=10000)
orange_now = probAboveGEV(max(tail(year_max$Value, n=25)), 1, 2023, 1960,
     estimates[1], (estimates[2]), estimates[3], cov=cov, samples=10000)
red_now = probAboveGEV(max(tail(year_max$Value, n=50)), 1, 2023, 1960,
    estimates[1], (estimates[2]), estimates[3], cov=cov, samples=10000)

return(list("name"= k,
            "loc_bm0"=  estimates[1],
            "low_loc_bm0"=  estimates[1]- 1.96*SError[1],
            "high_loc_bm0"=  estimates[1]+ 1.96*SError[1],
            "loc_se_bm0" = SError[1],
            "phi_bm0" = estimates[2],
            "low_phi_bm0"=  estimates[2]- 1.96*SError[2],
            "high_phi_bm0"=  estimates[2]+ 1.96*SError[2],
            "phi_se_bm0" = SError[2],
            "xi_bm0" = estimates[3],
            "low_xi_bm0"=  estimates[3]- 1.96*SError[3],
            "high_xi_bm0"=  estimates[3]+ 1.96*SError[3],
            "xi_tested_bm0" = test,
            "xi_se_bm0" = SError[3],
            "xi_p_bm0" = xi_p,
            "ad_bm0" = ad,
            "yellow_now_bm0" = yellow_now,
            "orange_now_bm0" = orange_now,
            "red_now_bm0" = red_now,
            "cov_11_bm0" = mat[1,1],
            "cov_21_bm0" = mat[2,1],
            "cov_31_bm0" = mat[3,1],
            "cov_41_bm0" = 0,
            "cov_51_bm0" = 0,
            "cov_61_bm0" = 0,
            "cov_22_bm0" = mat[2,2],
            "cov_32_bm0" = mat[3,2],
            "cov_42_bm0" = 0,
            "cov_52_bm0" = 0,
            "cov_62_bm0" = 0,
            "cov_33_bm0" = mat[3,3],
            "cov_43_bm0" = 0,
            "cov_53_bm0" = 0,
            "cov_63_bm0" = 0,
            "cov_44_bm0" = 0,
            "cov_54_bm0" = 0,
            "cov_64_bm0" = 0,
            "cov_55_bm0" = 0,
```

```
                "cov_65_bm0" = 0,
                "cov_66_bm0" = 0,
                "yellow_level" = max(tail(year_max$Value, n=5)),
                "orange_level" = max(tail(year_max$Value, n=25)),
                "red_level"=max(tail(year_max$Value, n=50))
  ))
}

# BM1 - Block maxima with trends in my
BM_my_function <- function(k,waterflow_data_transformad) {
  # Max per year
  year_max <- apply.yearly(waterflow_data_transformad, max)

  # The data needs to be in this form
  x <- matrix(ncol = 1, nrow = length(year_max$Value))
  x[,1] <- seq(1,length(year_max$Value),1)

  # ML estimate
  fit.gev <- gev.fit(year_max$Value, ydat= x, siglink= exp, mul = TRUE,
      show = FALSE, method="SANN")
  mat <- (fit.gev$cov)
  if (min(eigen(mat)$value) < 0){
    fit.gev <- gev.fit(year_max$Value, ydat= x, siglink= exp, mul = TRUE,
        show = FALSE, method="SANN", maxit = 1000000)
  }

  estimates <- fit.gev$mle
  SError <- fit.gev$se

  # Covariance matrix
  mat <- (fit.gev$cov)
  cov <- matrix(0,6,6)
  cov[1,1] = mat[1,1]
  cov[1,2] = cov[2,1] = mat[3,1]
  cov[3,1] = cov[1,3] = mat[4,1]
  cov[1,4] = cov[4,1] = mat[2,1]
  cov[2,2] = mat[3,3]
  cov[2,3] = cov[3,2] = mat[4,3]
  cov[2,4] = cov[4,2] = mat[3,2]
  cov[3,3] = mat[4,4]
  cov[3,4] = cov[4,3] = mat[4,2]
  cov[4,4] = mat[2,2]

  # Tests
  xi_test <- ifelse(0 > estimates[4]- 1.96*SError[4] & 0 < estimates[4]+
      1.96*SError[4], 0, estimates[4])
  loc1_p <- 2*(1-pnorm(abs(estimates[2]/SError[2]), 0, 1))
  xi_p <- 2*(1-pnorm(abs(estimates[4]/SError[4]), 0, 1))

  # Weather predictions
  yellow_30 = probAboveGEV(max(tail(year_max$Value, n=5)), 1, 2024+29,
      1960, estimates[1], (estimates[3]), estimates[4], loc1=estimates[2],
      phi1=0, shape1=0, cov=cov, samples=10000)
  orange_30 = probAboveGEV(max(tail(year_max$Value, n=25)), 1, 2024+29,
      1960, estimates[1], (estimates[3]), estimates[4], loc1=estimates[2],
      phi1=0, shape1=0, cov=cov, samples=10000)
  red_30 = probAboveGEV(max(tail(year_max$Value, n=50)), 1, 2024+29, 1960,
      estimates[1], (estimates[3]), estimates[4], loc1=estimates[2], phi1
      =0, shape1=0, cov=cov, samples=10000)

  yellow_now = probAboveGEV(max(tail(year_max$Value, n=5)), 1, 2023, 1960,
```

```
    estimates[1], (estimates[3]), estimates[4], loc1=estimates[2], phi1
    =0, shape1=0, cov=cov, samples=10000)
orange_now = probAboveGEV(max(tail(year_max$Value, n=25)), 1, 2023, 1960,
     estimates[1], (estimates[3]), estimates[4], loc1=estimates[2], phi1
    =0, shape1=0, cov=cov, samples=10000)
red_now = probAboveGEV(max(tail(year_max$Value, n=50)), 1, 2023, 1960,
    estimates[1], (estimates[3]), estimates[4], loc1=estimates[2], phi1
    =0, shape1=0, cov=cov, samples=10000)

yellow_ratio = yellow_30/yellow_now - 1
orange_ratio = orange_30/orange_now - 1
red_ratio = red_30/red_now - 1

return(list("name"= k,
            "loc0_bm1"=  estimates[1],
            "low_loc0_bm1"=  estimates[1]- 1.96*SError[1],
            "high_loc0_bm1"=  estimates[1]+ 1.96*SError[1],
            "loc0_se_bm1" = SError[1],
            "loc1_bm1" = estimates[2],
            "low_loc1_bm1"=  estimates[2]- 1.96*SError[2],
            "high_loc1_bm1"=  estimates[2]+ 1.96*SError[2],
            "loc1_se_bm1" = SError[2],
            "phi_bm1" = estimates[3],
            "low_phi_bm1"=  estimates[3]- 1.96*SError[3],
            "high_phi_bm1"=  estimates[3]+ 1.96*SError[3],
            "phi_se_bm1" = SError[3],
            "xi_bm1" = estimates[4],
            "low_xi_bm1"=  estimates[4]- 1.96*SError[4],
            "high_xi_bm1"=  estimates[4]+ 1.96*SError[4],
            "xi_tested_bm1" = xi_test,
            "xi_se_bm1" = SError[4],
            #"ad_bm1" = ad,
            "loc1_p_bm1" = loc1_p,
            "xi_p_bm1" = xi_p,
            "yellow_30_bm1" = yellow_30,
            "orange_30_bm1" = orange_30,
            "red_30_bm1" = red_30,
            "yellow_now_bm1" = yellow_now,
            "orange_now_bm1" = orange_now,
            "red_now_bm1" = red_now,
            "yellow_ratio_bm1" = yellow_ratio,
            "orange_ratio_bm1" = orange_ratio,
            "red_ratio_bm1" = red_ratio,
            "cov_11_bm1" = mat[1,1],
            "cov_21_bm1" = mat[3,1],
            "cov_31_bm1" = mat[4,1],
            "cov_41_bm1" = mat[2,1],
            "cov_51_bm1" = 0,
            "cov_61_bm1" = 0,
            "cov_22_bm1" = mat[3,3],
            "cov_32_bm1" = mat[4,3],
            "cov_42_bm1" = mat[3,2],
            "cov_52_bm1" = 0,
            "cov_62_bm1" = 0,
            "cov_33_bm1" = mat[4,4],
            "cov_43_bm1" = mat[4,2],
            "cov_53_bm1" = 0,
            "cov_63_bm1" = 0,
            "cov_44_bm1" = mat[2,2],
            "cov_54_bm1" = 0,
            "cov_64_bm1" = 0,
```

```
                "cov_55_bm1" = 0,
                "cov_65_bm1" = 0,
                "cov_66_bm1" = 0
  ))
}

# BM2 - Block maxima with trends in phi
BM_phi_function <- function(k,waterflow_data_transformad) {
  # Max per year
  year_max <- apply.yearly(waterflow_data_transformad, max)

  # The data needs to be in this form
  x <- matrix(ncol = 1, nrow = length(year_max$Value))
  x[,1] <- seq(1,length(year_max$Value),1)

  # ML estimate
  fit.gev_start_value <- gev.fit(year_max$Value, ydat= x, sigl = TRUE, show
      =FALSE, method="BFGS", maxit=100000)
  fit.gev <- gev.fit(year_max$Value, ydat= x, sigl = TRUE, siglink = exp,
      show=FALSE, method="BFGS", maxit=100000, muinit = fit.gev_start_value
      $mle[1], siginit = c(log(fit.gev_start_value$mle[2]), fit.gev_start_
      value$mle[3]/fit.gev_start_value$mle[2]))

  if (min(diag(fit.gev$cov)) < 0){
    fit.gev <- gev.fit(year_max$Value, ydat= x, sigl = TRUE, siglink = exp,
        show=FALSE, maxit=100000, muinit = fit.gev_start_value$mle[1],
      siginit = c(log(fit.gev_start_value$mle[2]), fit.gev_start_value$
      mle[3]/fit.gev_start_value$mle[2]))
  }

  if (min(diag(fit.gev$cov)) < 0){
    fit.gev <- gev.fit(year_max$Value, ydat= x, sigl = TRUE, siglink = exp,
        show=FALSE, method="SANN", maxit=100000, muinit = fit.gev_start_
      value$mle[1], siginit = c(log(fit.gev_start_value$mle[2]), fit.gev_
      start_value$mle[3]/fit.gev_start_value$mle[2]))
  }

  estimates <- fit.gev$mle
  SError <- fit.gev$se

  # Covariance matrix
  mat <- (fit.gev$cov)
  cov <- matrix(0,6,6)
  cov[1,1] = mat[1,1]
  cov[1,2] = cov[2,1] = mat[2,1]
  cov[3,1] = cov[1,3] = mat[4,1]
  cov[1,5] = cov[5,1] = mat[3,1]
  cov[2,2] = mat[2,2]
  cov[2,3] = cov[3,2] = mat[4,2]
  cov[2,5] = cov[5,2] = mat[3,2]
  cov[3,3] = mat[4,4]
  cov[3,5] = cov[5,3] = mat[4,3]
  cov[5,5] = mat[3,3]

  # Tests
  phi1_p <- 2*(1-pnorm(abs(estimates[3]/SError[3]), 0, 1))
  xi_p <- 2*(1-pnorm(abs(estimates[4]/SError[4]), 0, 1))

  # Weather predictions
  yellow_30 = probAboveGEV(max(tail(year_max$Value, n=5)), 1, 2024+29,
      1960, estimates[1], (estimates[2]), estimates[4], loc1=0, phi1=
```

```
    estimates[3], shape1=0, cov=cov, samples=10000)
orange_30 = probAboveGEV(max(tail(year_max$Value, n=25)), 1, 2024+29,
    1960, estimates[1], (estimates[2]), estimates[4], loc1=0, phi1=
    estimates[3], shape1=0, cov=cov, samples=10000)
red_30 = probAboveGEV(max(tail(year_max$Value, n=50)), 1, 2024+29, 1960,
    estimates[1], (estimates[2]), estimates[4], loc1=0, phi1=estimates
    [3], shape1=0, cov=cov, samples=10000)

yellow_now = probAboveGEV(max(tail(year_max$Value, n=5)), 1, 2023, 1960,
    estimates[1], (estimates[2]), estimates[4], loc1=0, phi1=estimates
    [3], shape1=0, cov=cov, samples=10000)
orange_now = probAboveGEV(max(tail(year_max$Value, n=25)), 1, 2023, 1960,
     estimates[1], (estimates[2]), estimates[4], loc1=0, phi1=estimates
    [3], shape1=0, cov=cov, samples=10000)
red_now = probAboveGEV(max(tail(year_max$Value, n=50)), 1, 2023, 1960,
    estimates[1], (estimates[2]), estimates[4], loc1=0, phi1=estimates
    [3], shape1=0, cov=cov, samples=10000)

yellow_ratio = yellow_30/yellow_now - 1
orange_ratio = orange_30/orange_now - 1
red_ratio = red_30/red_now - 1

return(list("name"= k,
            "loc_bm2"=  estimates[1],
            "low_loc_bm2"=  estimates[1]- 1.96*SError[1],
            "high_loc_bm2"=  estimates[1]+ 1.96*SError[1],
            "phi0_bm2" = (estimates[2]),
            "low_phi0_bm2"=  (estimates[2])- 1.96*(SError[2]),
            "high_phi0_bm2"=  (estimates[2])+ 1.96*(SError[2]),
            "phi1_bm2" = (estimates[3]),
            "low_phi1_bm2"=  (estimates[3])- 1.96*(SError[3]),
            "high_phi1_bm2"=  (estimates[3])+ 1.96*(SError[3]),
            "xi_bm2" = estimates[4],
            "low_xi_bm2"=  estimates[4]- 1.96*SError[4],
            "high_xi_bm2"=  estimates[4]+ 1.96*SError[4],
            #"xi_tested_bm2" = xi_test,
            #"ad_bm2" = ad,
            "phi1_p_bm2" = phi1_p,
            "xi_p_bm2" = xi_p,
            "loc_se_bm2" = SError[1],
            "phi0_se_bm2" = SError[2],
            "phi1_se_bm2" = SError[3],
            "xi_se_bm2" = SError[4],
            "yellow_30_bm2" = yellow_30,
            "orange_30_bm2" = orange_30,
            "red_30_bm2" = red_30,
            "yellow_now_bm2" = yellow_now,
            "orange_now_bm2" = orange_now,
            "red_now_bm2" = red_now,
            "yellow_ratio_bm2" = yellow_ratio,
            "orange_ratio_bm2" = orange_ratio,
            "red_ratio_bm2" = red_ratio,
            "cov_11_bm2" = mat[1,1],
            "cov_21_bm2" = mat[2,1],
            "cov_31_bm2" = mat[4,1],
            "cov_41_bm2" = 0,
            "cov_51_bm2" = mat[3,1],
            "cov_61_bm2" = 0,
            "cov_22_bm2" = mat[2,2],
            "cov_32_bm2" = mat[4,2],
            "cov_42_bm2" = 0,
```

```
                "cov_52_bm2" = mat[3,2],
                "cov_62_bm2" = 0,
                "cov_33_bm2" = mat[4,4],
                "cov_43_bm2" = 0,
                "cov_53_bm2" = mat[4,3],
                "cov_63_bm2" = 0,
                "cov_44_bm2" = 0,
                "cov_54_bm2" = 0,
                "cov_64_bm2" = 0,
                "cov_55_bm2" = mat[3,3],
                "cov_65_bm2" = 0,
                "cov_66_bm2" = 0

  ))
}

# PoT0 - Peaks over thereshold without trends
PoT_function <- function(k,waterflow_data_transformad) {
  # Max per year
  year_max <- apply.yearly(waterflow_data_transformad, max)

  # The data needs to be in this form
  x <- matrix(ncol = 1, nrow = length(waterflow_data_transformad$Value))
  x[,1] <- seq(1,length(waterflow_data_transformad$Value),1)

  # Decide peak
  u <- sort(waterflow_data_transformad$Value)[length(waterflow_data_
      transformad$Value)*0.98]
  declust <- c(decluster(waterflow_data_transformad$Value, u,replace.with=u
      -1))

  # Count clusters
  v=0
  for (i in declust){
    if (i > u){
      v <- v+1
    }
  }
  clusters <- v

  pre_lambda <- 365.25*(gpd.fit(waterflow_data_transformad$Value, threshold
       = u, ydat= x, siglink = exp, show = FALSE, method="BFGS")$rate)

  # ML estimate
  fit.gpd_start_value <- gpd.fit(declust, threshold = u, ydat= x, show =
      FALSE, method="BFGS", maxit = 100000)
  fit.gpd <- gpd.fit(declust, threshold = u, ydat= x, siglink = exp, show =
       FALSE, method="BFGS", maxit = 100000, siginit = log(fit.gpd_start_
      value$mle[1]))

  estimates <- fit.gpd$mle
  SError <- fit.gpd$se

  # Covariance matrix
  mat <- fit.gpd$cov
  cov = matrix(0,4,4)
  cov[1,1] = mat[1,1]
  cov[2,2] = mat[2,2]
  cov[1,2] = cov[2,1] = mat[2,1]

  # Make a vector with time between clusters
```

```
time <- 0
time_between_clusters <- c()
for (i in declust){
  if (i > u){
    time_between_clusters <- c(time_between_clusters, time)
  }
  time <- time +1
}

# Estimate lambdas
lambda <- 365.25*fit.gpd_start_value$rate
all.years <-seq(1,length(declust),length.out=length(declust))
out<-fitPP.fun(covariates=cbind(all.years), posE=time_between_clusters,
    start=list(b0=0,b1=0), modSim = TRUE)
omegaSE <- (sqrt(diag(out@vcov)))

# Tests
ad <- gnfit(waterflow_data_transformad$Value, "gpd", pr = c(exp(estimates
    [1]), estimates[2]), threshold = u)$Apval
xi_p <- 2*(1-pnorm(abs(estimates[2]/SError[2]), 0, 1))
omega_error <- 2*(1-pnorm(abs((out@detailsb$par[2]/omegaSE[2])), 0, 1))

# Weather predictions
yellow2=probAboveGP(level=max(tail(year_max$Value, n=5)), u=u, year=2025,
     zeroYear=1960, lambda=pre_lambda, phi0=estimates[1], shape0=
    estimates[2], cov=cov)
orange2=probAboveGP(level=max(tail(year_max$Value, n=25)), u=u, year
    =2025, zeroYear=1960, lambda=pre_lambda, phi0=estimates[1], shape0=
    estimates[2], cov=cov)
red2=probAboveGP(level=max(tail(year_max$Value, n=50)), u=u, year=2025,
    zeroYear=1960, lambda=pre_lambda, phi0=estimates[1], shape0=estimates
    [2], cov=cov)

yellow = probAboveDayGP(level=max(tail(year_max$Value, n=5)), u=u, year
    =2024, zeroYear=1960, lambda=pre_lambda, phi0=estimates[1], shape0=
    estimates[2], cov=cov)
orange=probAboveDayGP(level=max(tail(year_max$Value, n=25)), u=u, year
    =2024, zeroYear=1960, lambda=pre_lambda, phi0=estimates[1], shape0=
    estimates[2], cov=cov)
red=probAboveDayGP(level=max(tail(year_max$Value, n=50)), u=u, year=2024,
     zeroYear=1960, lambda=pre_lambda, phi0=estimates[1], shape0=
    estimates[2], cov=cov)

return(list("name"= k,
            "pre_lambda" = pre_lambda,
            "post_lambda" = lambda,
            "threshold_u" = u,
            "phi_pot0"=  estimates[1],
            "low_phi_pot0"=  estimates[1]- 1.96*SError[1],
            "high_phi_pot0"=  estimates[1]+ 1.96*SError[1],
            "phi_se_pot0" = SError[1],
            "xi_pot0" = estimates[2],
            "low_xi_pot0"=  estimates[2]- 1.96*SError[2],
            "high_xi_pot0"=  estimates[2]+ 1.96*SError[2],
            "xi_se_pot0" = SError[2],
            "xi_p_pot0" = xi_p,
            "ad_pot0" = ad,
            "lambda0" = out@detailsb$par[1] + log(365.25),
            "lambda1" = out@detailsb$par[2]*365.25,
            "lambda_cov_11" = out@vcov[1,1],
            "lambda_cov_22" = out@vcov[2,2]*365.25^2,
```

```
                "lambda_cov_21" = out@vcov[1,2]*365.25,
                "lambda0_se" = omegaSE[1],
                "lambda1_se" = omegaSE[2]*365.25,
                "lambda1_p" = omega_error,
                "post_datapts" = clusters,
                "yellow_tomorrow_pot0" = yellow,
                "orange_tomorrow_pot0" = orange,
                "red_tomorrow_pot0" = red,
                "yellow_next_pot0" = yellow2,
                "orange_next_pot0" = orange2,
                "red_next_pot0" = red2,
                "cov_11_pot0" = mat[1,1],
                "cov_21_pot0" = mat[2,1],
                "cov_31_pot0" = 0,
                "cov_41_pot0" = 0,
                "cov_22_pot0" = mat[2,2],
                "cov_32_pot0" = 0,
                "cov_42_pot0" = 0,
                "cov_33_pot0" = 0,
                "cov_43_pot0" = 0,
                "cov_44_pot0" = 0

  ))
}

# PoT2 - Peaks over thereshold with trends in phi
PoT_phi_function <- function(k,waterflow_data_transformad) {
  # max per year
  year_max <- apply.yearly(waterflow_data_transformad, max)

  # decide threshold
  u <- sort(waterflow_data_transformad$Value)[length(waterflow_data_
      transformad$Value)*0.98]
  declust <- c(decluster(waterflow_data_transformad$Value, u,replace.with=u
      -1))

  # count clusters
  v=0
  for (i in declust){
    if (i > u){
      v <- v+1
    }
  }
  clusters <- v

  # The data needs to be in this form
  x <- matrix(ncol = 1, nrow = length(declust))
  x[,1] <- seq(1,length(declust),1)

  # ML estimate, uses a modded version of gpd.fit that
  # uses numderive instead of optim for hessian
  fit.gpd_start_value <- gpd.fit_mod(waterflow_data_transformad$Value,
      threshold = u, ydat= x, sigl=TRUE, show = FALSE, method="BFGS", maxit
       = 30000)
  pre_lambda <- 365.25* (gpd.fit_mod(waterflow_data_transformad$Value,
      threshold = u, ydat= x, siglink = exp, sigl=TRUE, show = FALSE,
      method="BFGS", maxit = 30000, siginit = c(log(fit.gpd_start_value$mle
      [1]), fit.gpd_start_value$mle[2]/fit.gpd_start_value$mle[1]), shinit
      = fit.gpd_start_value$mle[3])$rate)

  fit.gpd_start_value <- gpd.fit_mod(declust, threshold = u, ydat= x, sigl=
```

```
    TRUE, show = FALSE, method="BFGS", maxit = 100000)
fit.gpd <- (gpd.fit_mod(declust, threshold = u, ydat= x, siglink = exp,
    sigl=TRUE, show = FALSE, maxit = 1000000, siginit = c(log(fit.gpd_
    start_value$mle[1]), fit.gpd_start_value$mle[2]/fit.gpd_start_value$
    mle[1])))

if (min(diag(fit.gpd$cov)) < 0){
  fit.gpd <- (gpd.fit_mod(declust, threshold = u, ydat= x, siglink = exp,
       sigl=TRUE, show = FALSE, maxit = 100000, siginit = c(log(fit.gpd_
      start_value$mle[1]), fit.gpd_start_value$mle[2]/fit.gpd_start_value
      $mle[1]) ))
}

if (min(diag(fit.gpd$cov)) < 0){
  fit.gpd <- (gpd.fit_mod(declust, threshold = u, ydat= x, siglink = exp,
       sigl=TRUE, method="SANN", show = FALSE, maxit = 100000, siginit =
      c(log(fit.gpd_start_value$mle[1]), fit.gpd_start_value$mle[2]/fit.
      gpd_start_value$mle[1])))
}

estimates <- fit.gpd$mle
SError <- fit.gpd$se

# covariance matrix
mat <- fit.gpd$cov
cov = matrix(0,4,4)
cov[1,1] = mat[1,1]
cov[2,2] = mat[3,3]
cov[1,2] = cov[2,1] = mat[3,1]
cov[1,3] = cov[3,1] = mat[2,1]*365.25
cov[3,3] = mat[2,2]*365.25^2
cov[2,3] = cov[3,2] = mat[3,2]*365.25

# tests
phi1_p <- 2*(1-pnorm(abs(estimates[2]/SError[2]), 0, 1))
xi_p <- 2*(1-pnorm(abs(estimates[3]/SError[3]), 0, 1))

# weather predictions
yellow2=probAboveGP(level=max(tail(year_max$Value, n=5)), u=u, year=2025,
     zeroYear=1960, lambda=pre_lambda, phi0=estimates[1], shape0=
    estimates[3], phi1=estimates[2]*365.25, cov=cov)
orange2=probAboveGP(level=max(tail(year_max$Value, n=25)), u=u, year
    =2025, zeroYear=1960,lambda=pre_lambda, phi0=estimates[1], shape0=
    estimates[3], phi1=estimates[2]*365.25, cov=cov)
red2=probAboveGP(level=max(tail(year_max$Value, n=50)), u=u, year=2025,
    zeroYear=1960, lambda=pre_lambda, phi0=estimates[1], shape0=estimates
    [3], phi1=estimates[2]*365.25, cov=cov)

yellow=probAboveDayGP(level=max(tail(year_max$Value, n=5)), u=u, year
    =2024, zeroYear=1960, lambda=pre_lambda, phi0=estimates[1], shape0=
    estimates[3], phi1=estimates[2]*365.25, cov=cov)
orange=probAboveDayGP(level=max(tail(year_max$Value, n=25)), u=u, year
    =2024, zeroYear=1960, lambda=pre_lambda, phi0=estimates[1], shape0=
    estimates[3], phi1=estimates[2]*365.25, cov=cov)
red=probAboveDayGP(level=max(tail(year_max$Value, n=50)), u=u, year=2024,
     zeroYear=1960, lambda=pre_lambda, phi0=estimates[1], shape0=
    estimates[3], phi1=estimates[2]*365.25, cov=cov)

return(list("name"= k,
            "phi0_pot2"=  estimates[1],
            "low_phi0_pot2"=  estimates[1]- 1.96*SError[1],
```

```
                "high_phi0_pot2"=  estimates[1]+ 1.96*SError[1],
                "phi0_se_pot2" = SError[1],
                "phi1_pot2"=  estimates[2]*365.25,
                "low_phi1_pot2"=  365.25*(estimates[2]- 1.96*SError[2]),
                "high_phi1_pot2"=  365.25*(estimates[2]+ 1.96*SError[2]),
                "phi1_se_pot2" = SError[2]*365.25,
                "xi_pot2" = estimates[3],
                "low_xi_pot2"=  estimates[3]- 1.96*SError[3],
                "high_xi_pot2"=  estimates[3]+ 1.96*SError[3],
                "xi_se_pot2" = SError[3],
                "xi_p_pot2" = xi_p,
                "yellow_tomorrow_pot2" =  yellow,
                "orange_tomorrow_pot2" = orange,
                "red_tomorrow_pot2" = red,
                "yellow_next_pot2" =  yellow2,
                "orange_next_pot2" = orange2,
                "red_next_pot2" = red2,
                "phi1_p_pot2" = phi1_p,
                "cov_11_pot2" = cov[1,1],
                "cov_21_pot2" = cov[2,1],
                "cov_31_pot2" = cov[3,1],
                "cov_41_pot2" = 0,
                "cov_22_pot2" = cov[2,2],
                "cov_32_pot2" = cov[3,2],
                "cov_42_pot2" = 0,
                "cov_33_pot2" = cov[2,2],
                "cov_43_pot2" = 0,
                "cov_44_pot2" = 0
  ))
}

# Function for Benjamini-Hochberg
BH_function <- function(data, column, k){
  BH <- list((data["name"]), p.adjust(t(data[column]), method = "BH"))
  names(BH) <- c("name", k)
  data = merge(data, data.frame(BH), by="name")
  return(data)
}

# Tests if the p-values are significant
tester <- function(data, column, replacer, k){
  c <-c()
  for (i in 1:length(data[,column])){
    if (is.nan(data[column][i,1])){
      data[column][i,1] <- 1
    }
    if (data[column][i,1] < 0.05){
      c <-c(c, data[replacer][i,1])
    }
    else{
      c <- c(c, 0)
    }
  }
  ans <- list(data["name"], c)
  names(ans) <- c("name", k)
  return(merge(data, data.frame(ans), by = "name"))
}

list <- list.files("data")

bm0 = NULL
```

```
bm1 = NULL
bm2 = NULL
pot0 = NULL
pot2 = NULL

# Loops through all files
for (i in list) {
  # Read data from CSV
  path <- c("data/","")
  data <- read.csv(paste(path, collapse=i), sep = ";")

  # Rewrites the data into date and value form
  waterflow_data_transformad <- data.frame(
    Date = as.Date(data$Datum),
    Value = data$Vatten
  )

  waterflow_data_transformad$Date <- waterflow_data_transformad$Date %m+%
      months(9)

  # Run the models
  bm0 = rbind(bm0, data.frame(BM_function(i, waterflow_data_transformad)))
  bm1 = rbind(bm1, data.frame(BM_my_function(i, waterflow_data_transformad)
      ))
  bm2 = rbind(bm2, data.frame(BM_phi_function(i, waterflow_data_transformad
      )))
  pot0 = rbind(pot0, data.frame(PoT_function(i, waterflow_data_transformad)
      ))
  pot2 = rbind(pot2, data.frame(PoT_phi_function(i, waterflow_data_
      transformad)))
}

# Benjamini-Hochberg for Anderson-Darling and null tests
bm0 = BH_function(bm0, "ad_bm0", "adbh_bm0")
bm0 = BH_function(bm0, "xi_p_bm0", "xi_bh_bm0")

bm1 = BH_function(bm1, "loc1_p_bm1", "loc1_bh_bm1")
bm1 = BH_function(bm1, "xi_p_bm1", "xi_bh_bm1")

bm2 = BH_function(bm2, "phi1_p_bm2", "phi1_bh_bm2")
bm2 = BH_function(bm2, "xi_p_bm2", "xi_bh_bm2")

pot0 = BH_function(pot0, "ad_pot0", "adbh_pot0")
pot0 = BH_function(pot0, "xi_p_pot0", "xi_bh_pot0")
pot0 = BH_function(pot0, "lambda1_p", "lambda1_bh")

pot2 = BH_function(pot2, "phi1_p_pot2", "phi1_bh_pot2")
pot2 = BH_function(pot2, "xi_p_pot2", "xi_bh_pot2")

# Checks if the trend parameters are zero
bm1 = tester(bm1, "loc1_bh_bm1","loc1_bm1","loc1_tested_bm1")
bm2 = tester(bm2, "phi1_bh_bm2","phi1_bm2","phi1_tested_bm2")

# Merge all models
Merged_data <- merge(bm0,bm1,by="name")
Merged_data <- merge(Merged_data,bm2,by="name")
Merged_data <- merge(Merged_data,pot0,by="name")
Merged_data <- merge(Merged_data,pot2,by="name")

# Write the csv
write.csv2(Merged_data, "All␣data.csv", row.names=FALSE)
```

C.4 Exempel av plottgenerering i Python

Nedan visas hur prediktionsnivåplottarna har genererats.

```
# Edvin
import pandas as pd
import os
import numpy as np
import matplotlib.pyplot as plt
from collections import namedtuple
import matplotlib.colors as mcolors
from matplotlib.cm import ScalarMappable

plt.rcParams["svg.fonttype"] = "none"
River = namedtuple("River", ["name", "data"])

plot_types = ["pred", "proj"]
methods = ["bm0", "bm1", "bm2"]

for plot_type in plot_types:
    for method in methods:
        savepath = r"Python/Grafresultat_tex/ProjPredplottar"
        river_directory = r"R/CSVs"
        river_directory = os.path.join(river_directory, plot_type)
        river_directory = os.path.join(river_directory, method)
        savepath = os.path.join(savepath, plot_type)
        savepath = os.path.join(savepath, method)
        rivers = []
        for filename in os.listdir(river_directory):
            filepath = os.path.join(river_directory, filename)
            if not os.path.isfile(filepath):
                continue
            river_name = filename.split(".")[0]
            rivers.append(
                River(
                    name=river_name,
                    data=pd.read_csv(
                        filepath,
                        delimiter=",",
                        decimal=".",
                    ),
                )
            )
        low_column_char = "_h"
        high_column_char = "_l"
        low_column_names = [
            str(round(i, 2)) + low_column_char for i in list(np.arange
                (0.01, 1, 0.01))
        ]
        high_column_names = [
            str(round(i, 2)) + high_column_char for i in list(np.arange
                (0.01, 1, 0.01))
        ]
        column_names = low_column_names + high_column_names

        for river in rivers:
            print(f"plotting␣{river.name},␣{plot_type},␣{method}")
            cmap = plt.get_cmap("plasma")
            norm = mcolors.Normalize(vmin=0, vmax=len(column_names) / 2)

            river.data["Unnamed:␣0"] = river.data["Unnamed:␣0"] + 2024
```

```
for i in range(len(high_column_names) - 1):
    alpha = 1.0 / len(column_names)
    color = cmap(norm(i))
    fill_between = True
    if fill_between:
        plt.fill_between(
            river.data["Unnamed: 0"],
            river.data[low_column_names[i]],
            river.data[low_column_names[i + 1]],
            color=color,
        )
        plt.fill_between(
            river.data["Unnamed: 0"],
            river.data[high_column_names[i]],
            river.data[high_column_names[i + 1]],
            color=color,
        )

plot_lines = True
if plot_lines:
    plt.plot(
        river.data["Unnamed: 0"],
        river.data["0.95_l"],
        color="black",
        linestyle="dashed",
        linewidth=1,
        label=r"\nittiofemprocentprediktionsintervall",
    )
    plt.plot(
        river.data["Unnamed: 0"],
        river.data["0.95_h"],
        color="black",
        linestyle="dashed",
        linewidth=1,
    )
    plt.plot(
        river.data["Unnamed: 0"],
        river.data["mean"],
        color="grey",
        linewidth=1,
        label="Vantevarde",
    )

savepath_for_file = os.path.join(
    savepath, f"{river.name}_{plot_type}_{method}.svg"
)
plt.tick_params(axis="x", which="major", pad=12)
plt.tick_params(axis="y", which="major", pad=5)
plt.xticks([2025, 2035, 2045, 2055])

plt.ylabel(r"\ylabel", labelpad=15)
plt.xlabel("Ar", labelpad=15)
plt.legend(borderpad=1, framealpha=0, fontsize=15)

sm = ScalarMappable(cmap=cmap, norm=norm)
sm.set_array(
    []
)

ticks =[1,20, 40, 60, 80, 99]
cbar = plt.colorbar(sm, ax=plt.gca(), ticks=ticks)
```

```
cbar.set_label(r"\konfidensintervallprocent")
plt.savefig(savepath_for_file, bbox_inches="tight")
plt.close()
```