Théorie de Valeurs Extrêmes

Pluies Maximales Jounalières à Marseille

Haeji Yun

2023-08-26

Dans ce projet, nous cherchons à estimer le niveau de retour de pluie journalière à Marseille, c'est-à-dire le niveau extrême de pluie que nous attendons à dépasser dans 100 ans et dans 1000 ans.

Pour cela, nous allons étudier un jeu de données qui contient l'accumulation de pluie journalière en $10^{-1}mm$ à Marseille pendant 127 ans depuis le 1er août 1864 jusqu'au 31 juillet 1991.

Le jeu de données est sous forme d'un vecteur de dimension 46.355 qui correspond à 365 jours x 127 ans et il n'a pas de valeurs manquantes.

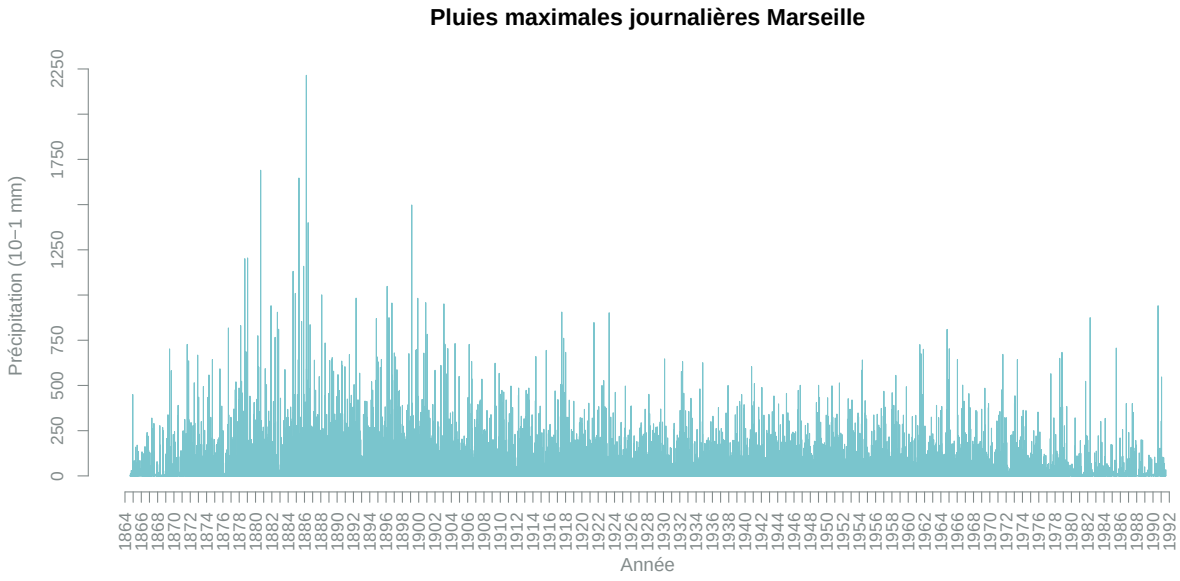
```
## [1] "Dimension: 46355"
```

```
## [1] "Valeurs manquantes: 0"
```

1. Étude Préliminaire

En convertissant le jeu de données en type séries temporelles, nous pouvons le visualiser en chronogramme pour avoir un aperçu global de données.

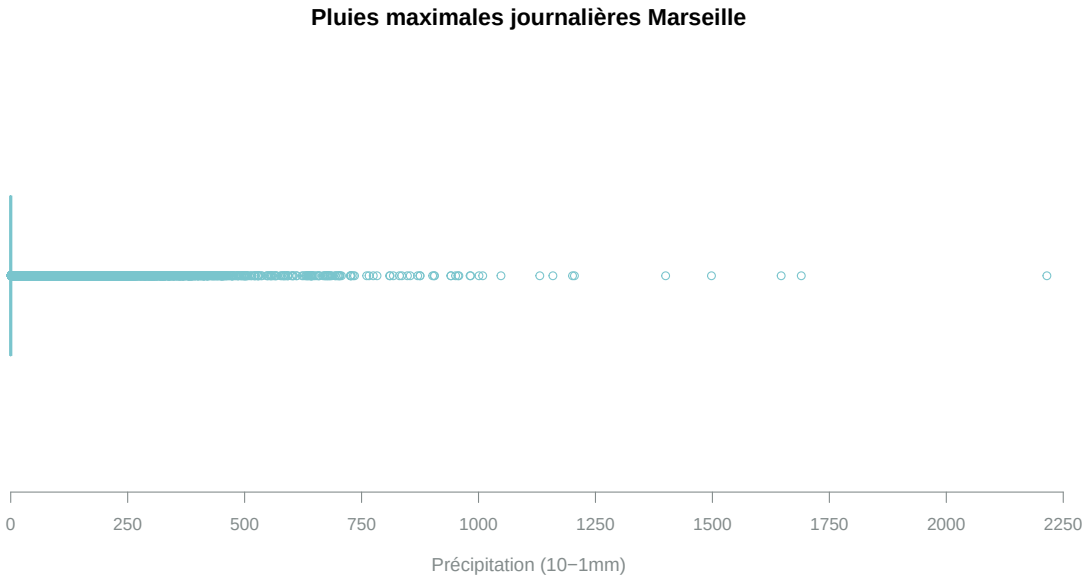
Nous obsesrvons une grande variation dans l'année et entre les différentes années. Dans la plupart des années, le niveau maximal journalier de pluies reste en dessous de $750^{-1}mm$, voire $500^{-1}mm$ et nous observons des niveaux particulièrement élevés entre la fin des années 70s aux années 80s.



Le niveau de pluie varie entre $0mm$ et $2.215^{-1}mm$. Avec les quartiles, nous remarquons également qu'il n'a pas plu à Marseille dans plus de 75% du temps.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00    0.00   15.84    0.00 2215.00
```

Pour comprendre mieux les jours de pluie, nous pouvons utiliser le boxplot. Nous obsesrvons que la pluie n'est pas un événement courant à Marseille et le niveau de pluie reste en générale en dessous de $750^{-1}mm$. Surtout, il y a eu que 5 jours parmi les 46.355 jours où il a plu plus de $1.250^{-1}mm$ et le niveau maximal observé de $2.250^{-1}mm$ s'écarte extrêmement des autres.



2. Loi d'extremum généralisée

Approche de maxima par blocs

Puisque nous cherchons le niveau extrême de pluies, ce sont les maxima que nous voudrons estimer. Pour cela, nous allons extraire les maxima de notre jeu de données pour en obtenir un échantillon.

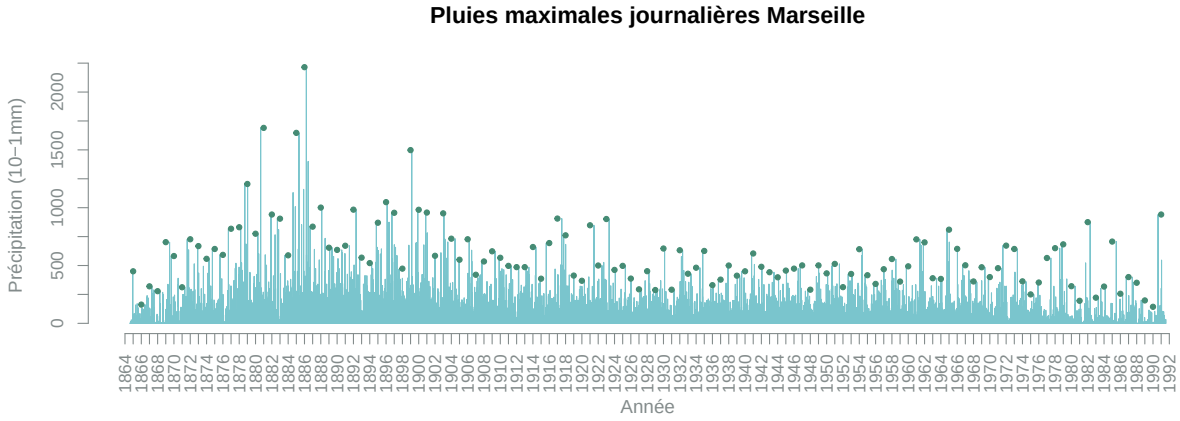
Nous pouvons utiliser l'approche de maxima par blocs où chaque bloc correspond à une année et on y extrait la valeur maximale de l'année. Nous aurons ainsi 127 valeurs comprises entre 143 et 2.215.

Ici, nous ne nous intéressons plus à que des maxima. En fait, si nous agrandissons le bloc pour avoir plusieurs années dans chaque bloc, nous n'aurons pas assez de données dans l'échantillon. Au contraire, si nous réduisons le bloc aux mois, nous aurons plus de données mais nous pouvons nous retrouver avec des petites valeurs et nous ne serons plus dans la queue de distribution, qui est la partie qui nous intéresse. Il y a donc un compromis biais - variance. En générale, pour éviter la saisonnalité des mois, le bloc annuel est considéré comme une bonne approche.

Dans notre échantillon obtenue, il n'y a pas eu d'années où il n'a pas plu. Néanmoins, nous remarquons qu'il contient également un niveau de pluie relativement bas.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   143.0   400.0   514.0   594.6   701.0  2215.0
```

Regardons les valeurs extraites sur le chronogramme. Les ronds verts correspondent à des valeurs maximales annuelles. À cause de grande variabilité entre les années, nous remarquons qu'une partie des grandes valeurs intéressantes est omise de notre échantillon.



Sur notre échantillon de maxima, nous pouvons ajuster la loi de valeurs extrêmes généralisée pour estimer les paramètres de la loi par la méthode du maximum de vraisemblance et la méthode de moments.

Estimation par le maximum de vraisemblance

Estimons les paramètres par la méthode du maximum de vraisemblance. Les paramètres estimés sont le paramètre de position b, le paramètre d'échelle a, et le paramètre de forme γ. Nous allons particulièrement nous intéresser au paramètre de forme car c'est γ qui nous donne l'information sur le comportement de loi.

L'estimation par le maximum de vraisemblance nous donne un γ positif, égal à 0,10. Nous avons alors un domaine d'attraction *Fréchet* avec un comportement convexe des valeurs extrêmes.

```
##
## fevd(x = maxPluies, type = "GEV", method = "MLE")
##
## [1] "Estimation Method used: MLE"
##
##
##  Negative Log-Likelihood Value:  878.1427
##
##
##  Estimated parameters:
##    location       scale       shape
## 459.3143744 196.4665623   0.1002232
##
##  Standard Error Estimates:
##    location       scale       shape
## 19.43883927 14.47810031  0.06065428
##
##  Estimated parameter covariance matrix.
##             location        scale        shape
## location 377.8684723 131.20968614 -0.346886744
## scale    131.2096861 209.61538849 -0.093006217
## shape     -0.3468867  -0.09300622  0.003678941
##
```

```
##  AIC = 1762.285
##
##  BIC = 1770.818
```

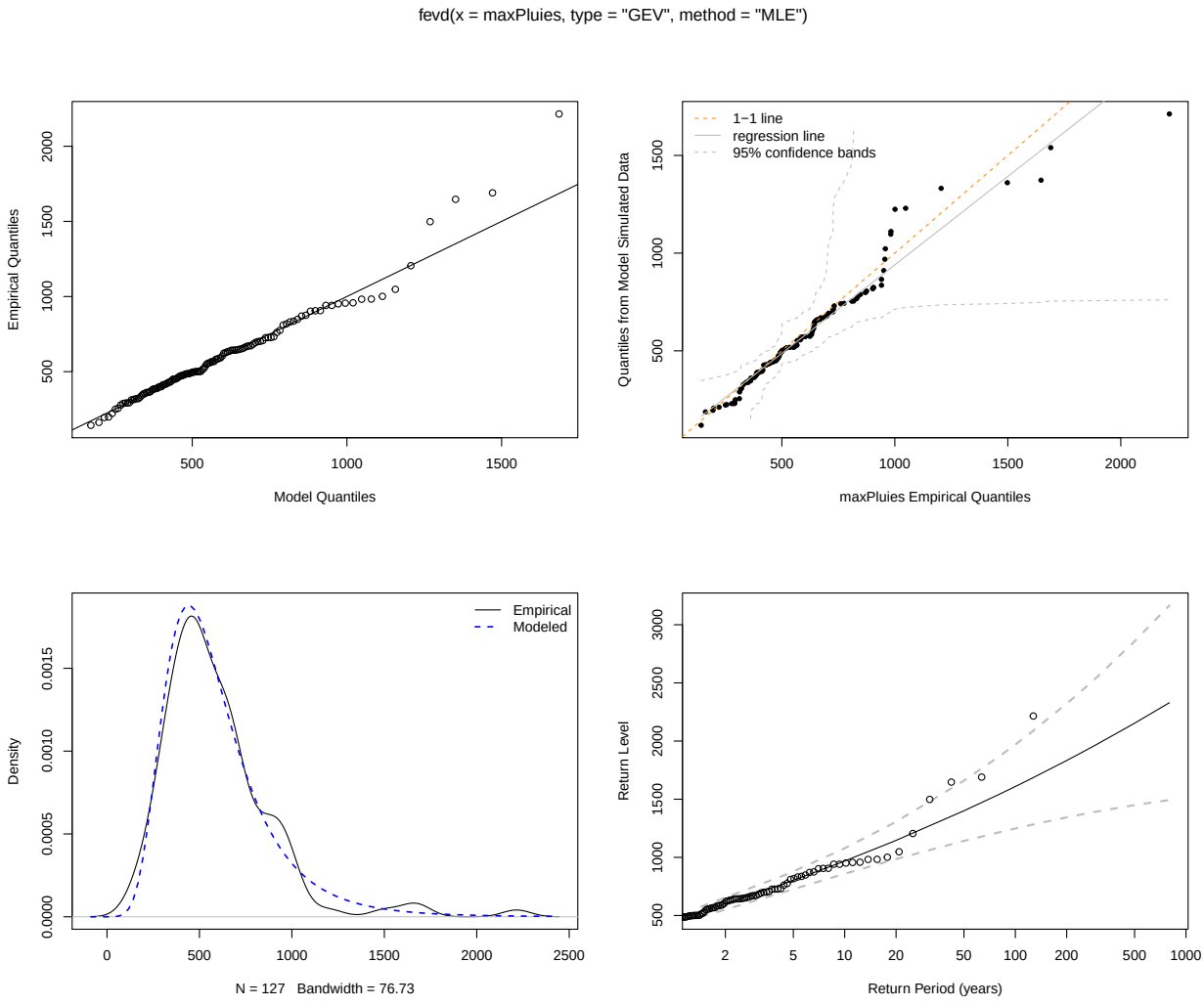
Nous pouvons anlayser graphiquement la qualité d'ajustement :

Le premier graphique représente le quantile empirique par rapport au quantile du modèle. Si les données observées correspondent bien aux données du modèle, elles doivent être alignées sur la droite. Nous observons que les données s'ajustent bien sur la droite avec quelques points qui s'éloigent de la droite. Il y a une sous-estimation des grandes valeurs à droite.

Le deuxième graphique nous permet de comparer les deux quantiles précédants avec la droite de régression en plus. Nous observons un bon ajustement mais toujours une sous-estimation à droite.

Le troisième graphique nous donne la densité empirique et la densité du modèle. Les deux densités se correspondent bien dans la globalité malgré un écart que nous observons vers la queue de la distribution, à droite. L'ajustement semble bon dans la globalité.

Le quatrième graphique représente le niveau de retour avec son intervalle de confiance par rapport à la période de retour. Si l'ajustement est bon, les données doivent être alignées sur la droite et se retrouver dans l'intervalle de confiance. Ici, nous avons un ajustement plutôt bien mais il y a 3 observations qui sortent de l'intervalle de confiance.



Pour les périodes de retour de 100 ans et de 1000 ans, le niveau de retour estimé est chacun 1.607 et 2.416. Sachant que la valeur maximale sur 127 ans soit 2.215, les valeurs estimées semblent basses.

```
## fevd(x = maxPluies, type = "GEV", method = "MLE")
##
## [1] "Normal Approx."
##
##                         95% lower CI Estimate 95% upper CI
## 100-year return level      1247.458 1607.505     1967.553
## 150-year return level      1305.202 1736.971     2168.740
## 1000-year return level     1510.947 2416.156     3321.366
```

Sur l'intervalle de confiance de paramètres, nous observons qu'il y a une incertitude. En effet, le calcul de l'intervalle de confiance nous montre qu'il y la probabilité que γ sois nul, d'où la probabilité d'être dans le domaine d'attraction de *Gumbel*.

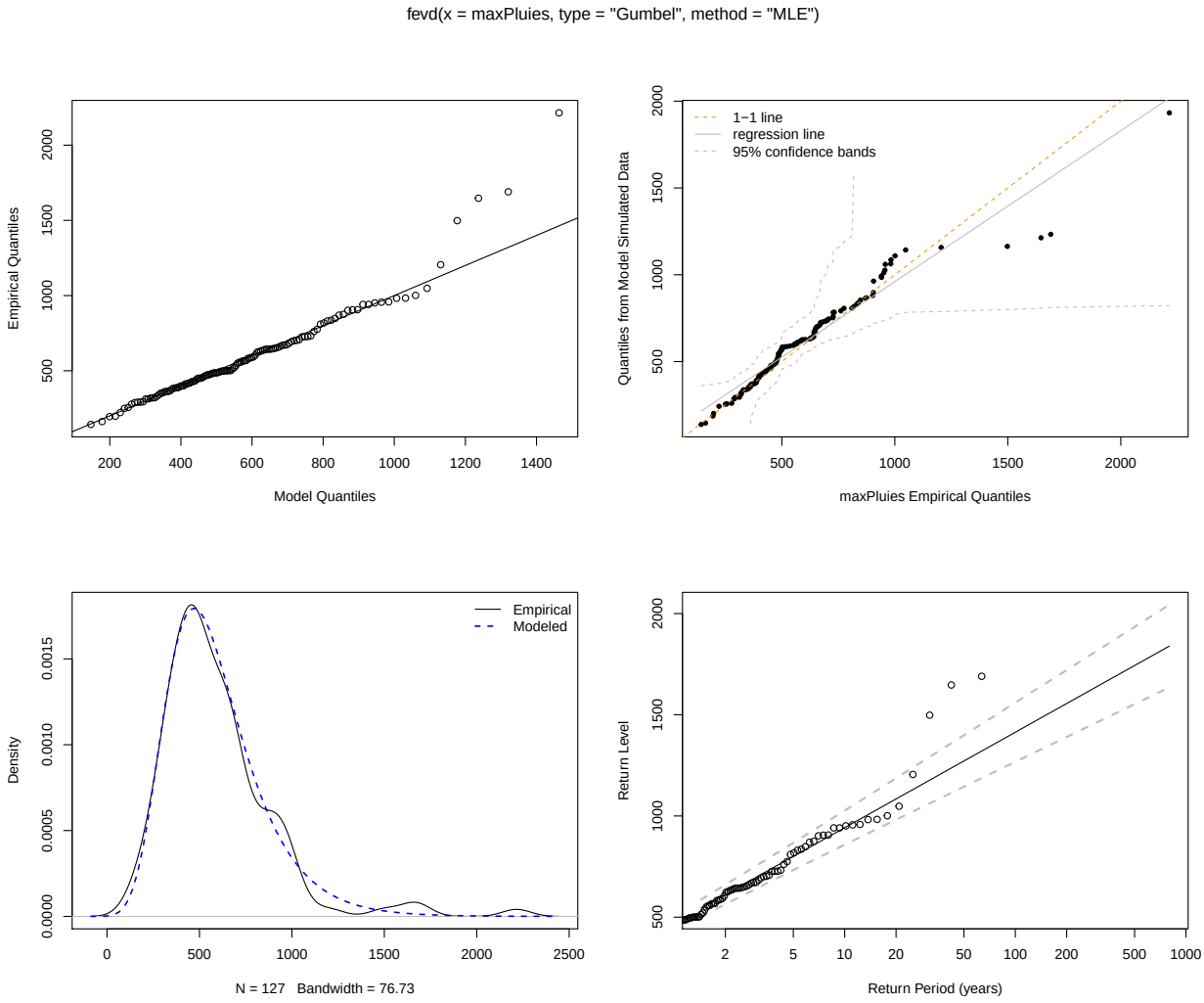
```
## fevd(x = maxPluies, type = "GEV", method = "MLE")
##
## [1] "Normal Approx."
##
##          95% lower CI    Estimate 95% upper CI
## location   421.214950 459.3143744  497.4137993
## scale      168.090007 196.4665623  224.8431175
## shape       -0.018657   0.1002232    0.2191034
```

Nous pouvons alors ajuster selon la loi de *Gumbel* pour vérifier si ceci est plausible.

```
##
## fevd(x = maxPluies, type = "Gumbel", method = "MLE")
##
## [1] "Estimation Method used: MLE"
##
##
##  Negative Log-Likelihood Value:  879.8518
##
##
##  Estimated parameters:
## location    scale
## 470.3556 204.8343
##
##  Standard Error Estimates:
## location    scale
## 19.04681 14.48100
##
##  Estimated parameter covariance matrix.
##           location     scale
## location 362.78105  82.05967
## scale     82.05967 209.69926
##
##  AIC = 1763.704
##
##  BIC = 1769.392
```

Sur les graphiques obtenus avec le modèle *Gumbel*, nous obtenons des résultats proches à ceux de la loi de *Fréchet*. L'ajustement est bon dans la globalité avec une sous-estimation pour les grandes valeurs à droite. Néanmoins sur le quatrième graphique, les valeurs hors intervalle de confiance est beaucoup plus éloignées de l'intervalle. Ceci est dû au fait que l'intervalle de confiance est droit et est plus restreint par rapport à la loi de *Fréchet*

Le modèle *Fréchet* est préférable au modèle *Gumbel*.



Le modèle *Gumbel* estime le niveau de retour égale à 1.412 pour la période de retour de 100 ans qui est encore plus bas que celui estimé par le modèle *Fréchet*. Ce niveau de retour est également inférieur à notre valeur maximale sur 127 ans. L'extrapolation n'est pas idéale avec le modèle *Gumbel*

```
## fevd(x = maxPluies, type = "Gumbel", method = "MLE")
##
## [1] "Normal Approx."
##
## [1] "100-year return level: 1412.624"
##
## [1] "95% Confidence Interval: (1266.5406, 1558.7075)"
```

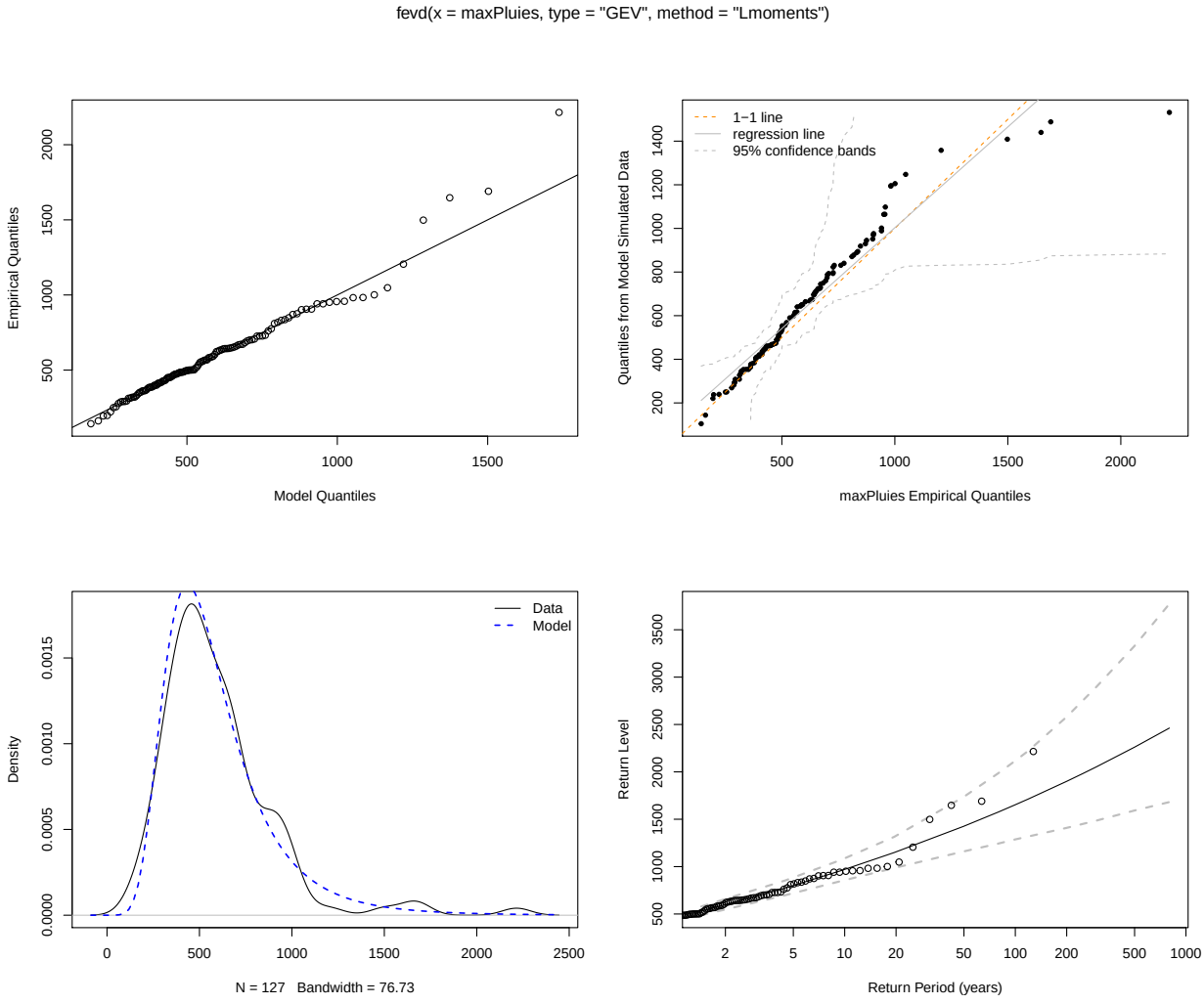
Estimation par la méthode des moments

Nous pouvons également estimer les paramètres avec la méthode de moments. Nous retrouvons un γ positif, égal à 0,12, qui est proche à l'estimation obtenue par le maximum de vraisemblance.

```
## fevd(x = maxPluies, type = "GEV", method = "Lmoments")
## [1] "GEV  Fitted to  maxPluies  using L-moments estimation."
##    location       scale       shape
## 456.7096967 192.5624483   0.1241941
```

Sur les graphiques, nous pouvons observer un bon ajustement. Malgré une sous-estimation des grandes valeurs à droite, les données sont bien alignées sur les droites quantile-quantile. L'ajustement est bon sur la densité. Et cette fois-ci, nous avons tous les données qui se trouvent dans l'intervalle de confiance sur le return level plot.

Vu le petit nombre d'observations que nous avons, la méthode par moment semble ajuster mieux que la méthode par le maximum de vraisemblance.



Le modèle ajusté par la méthode du moment estime le niveau de retour égal à 1.651 et 2.562 pour les périodes de retour de 100 ans et de 1000 ans respectivement. Bien qu'il estime des valeurs légèrement plus élevée, l'estimation semble toujours basse.

```
## fevd(x = maxPluies, type = "GEV", method = "Lmoments")
##
## [1] "Parametric Bootstrap"
## 502  iterations
##
##               2.5% Estimate    97.5%
## 100-year  1313.132 1651.503 2123.273
## 150-year  1393.517 1793.887 2373.290
## 1000-year 1736.863 2562.380 3982.186
```

3. Loi de pareto généralisée

Méthode de seuil

La loi de pareto généralisée modélise les excès.

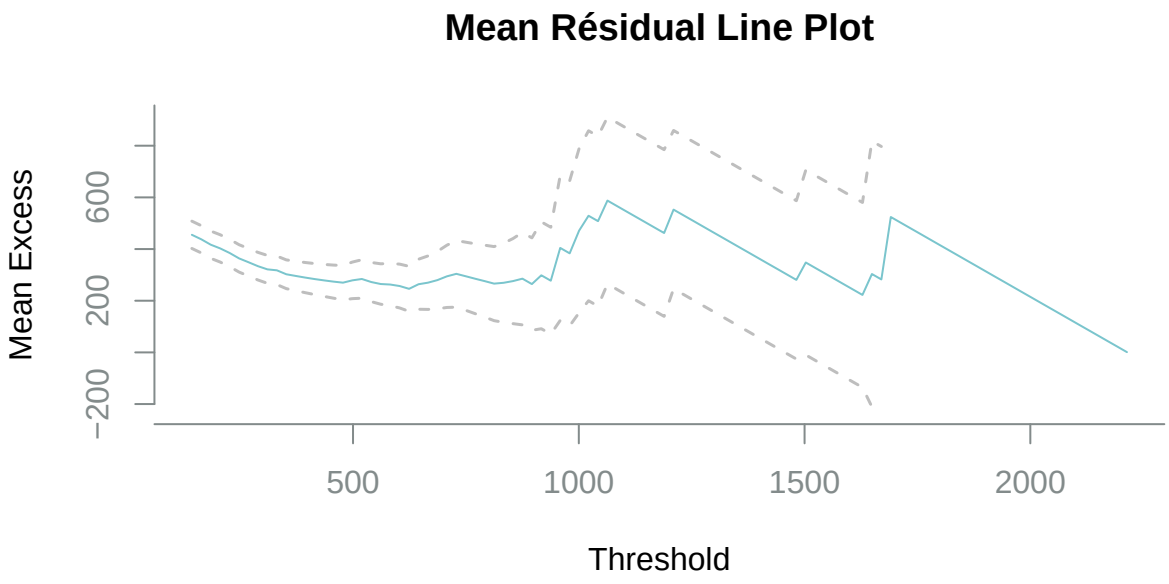
Les excès sont définis comme les valeurs qui dépassent un certin seuil moins la valeur du seuil. Pour ajuster selon la loi de pareto généralisée, nous avons besoin d'un échantillon des excès. Et nous voudrons que le nombre d'exèces dans notre échantillon soit plus élevé que le nombre de données dans l'échantillon obtenu par l'approche de maxima par bloc. Pour cela, le seuil doit être inférieur à 504.

```
## [1] "Seuil pour avoir plus de 127 données: 504"
```

Le choix de seuil demande également un compromis biais-variance. Si nous choisissons un seuil trop bas, nous ne sommes plus dans la queue de la distribution donc nous aurons un grand biais. Si nous fixons le seuil trop haut, nous aurons peu de données donc forte variance avec une grande incertitude.

Nous pouvons nous baser sur le mean residual line plot pour avoir une tranche de seuils plausibles. Nous considérons la valeur à partir de laquelle le plot n'est plus linéaire comme le seuil.

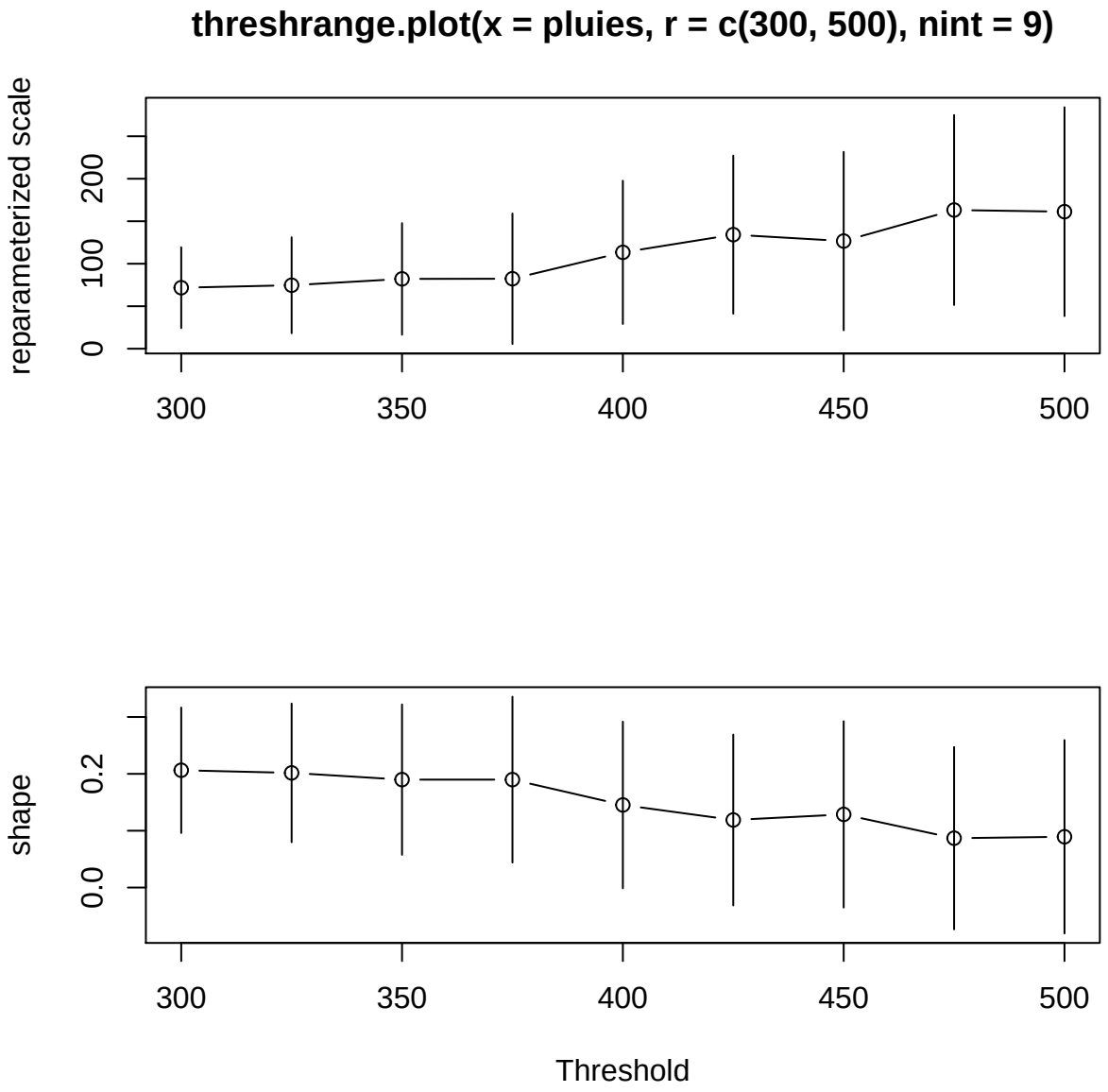
Dans notre graphique, nous avons une tendance linéaire jusq'à l'entour de 500 à 600. Comme nous voudrons un seuil inférieur à 504, nous allons nous intéresser aux valeurs inférieures à 500 et nous allons tester celles de la partie du graphique ayant la tendance linéaire, soit la tranche 300 à 500.



En effet, pour une loi d'excès donnée, le γ reste le même quelque soit l'excès du seuil, donc nous pouvons choisir le seuil le plus bas possible pour avoir plus de données dans l'échantillon.

Pour cela, nous pouvons regarder la stabilité de l'estimation de paramètre pour les différents seuils plausibles. Nous pouvons choisir la valeur jusqu'à laquelle l'estimation reste stable.

Entre l'intervalle 300 et 500, l'estimation reste stable jusqu'à 350.



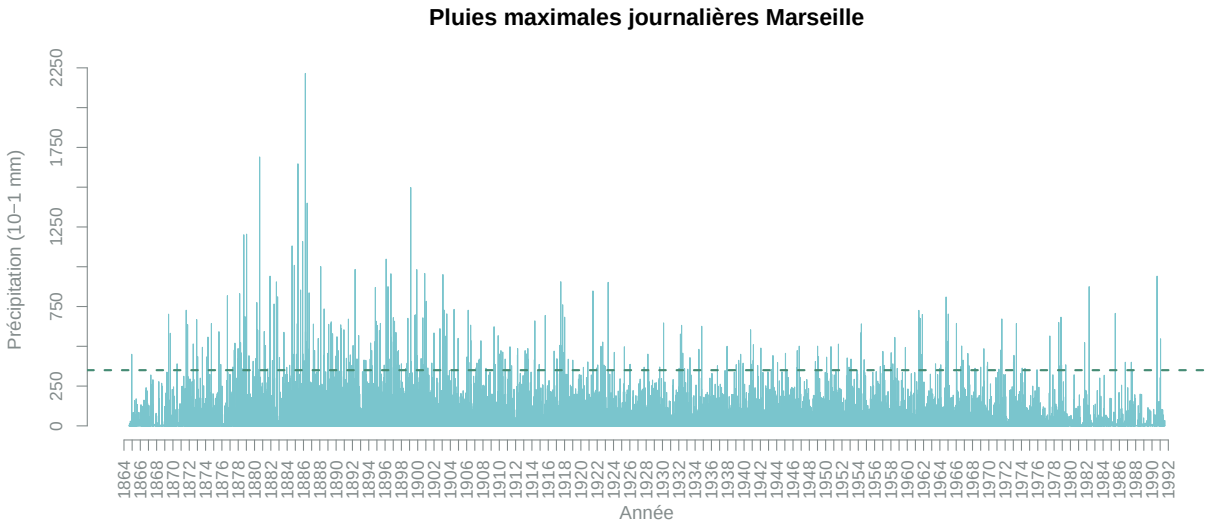
Avec un seuil de 350, nous avons 338 données qui représente 0,01% de nos données originales. On a beaucoup plus de données avec plus de valeurs qui sont issues de la queue de la distribution.

Nous avons un gain par rapport au cas précédant puisque nous avons perdu moins de données en gardant plus d'information pertinante.

```
## [1] "Dimension : 338"
```

```
## [1] "Quantile : 0.992708445690864"
```

En fixant le seuil à 350, nous gardons tous les valeurs qui se trouvent au dessus de 350. Ces valeurs seront notre echantillon d'excès.



Estimation par le maximum de vraisemblance

Nous pouvons ajuster la loi des excès directement sur le jeu de données original en précisant le seuil. Nous obtenons deux paramètres, le paramètre d'échelle σ et le paramètre de forme γ. Là aussi, nous nous intéressons sur le paramètre de forme qui nous donne l'indication sur le comportement des excès.

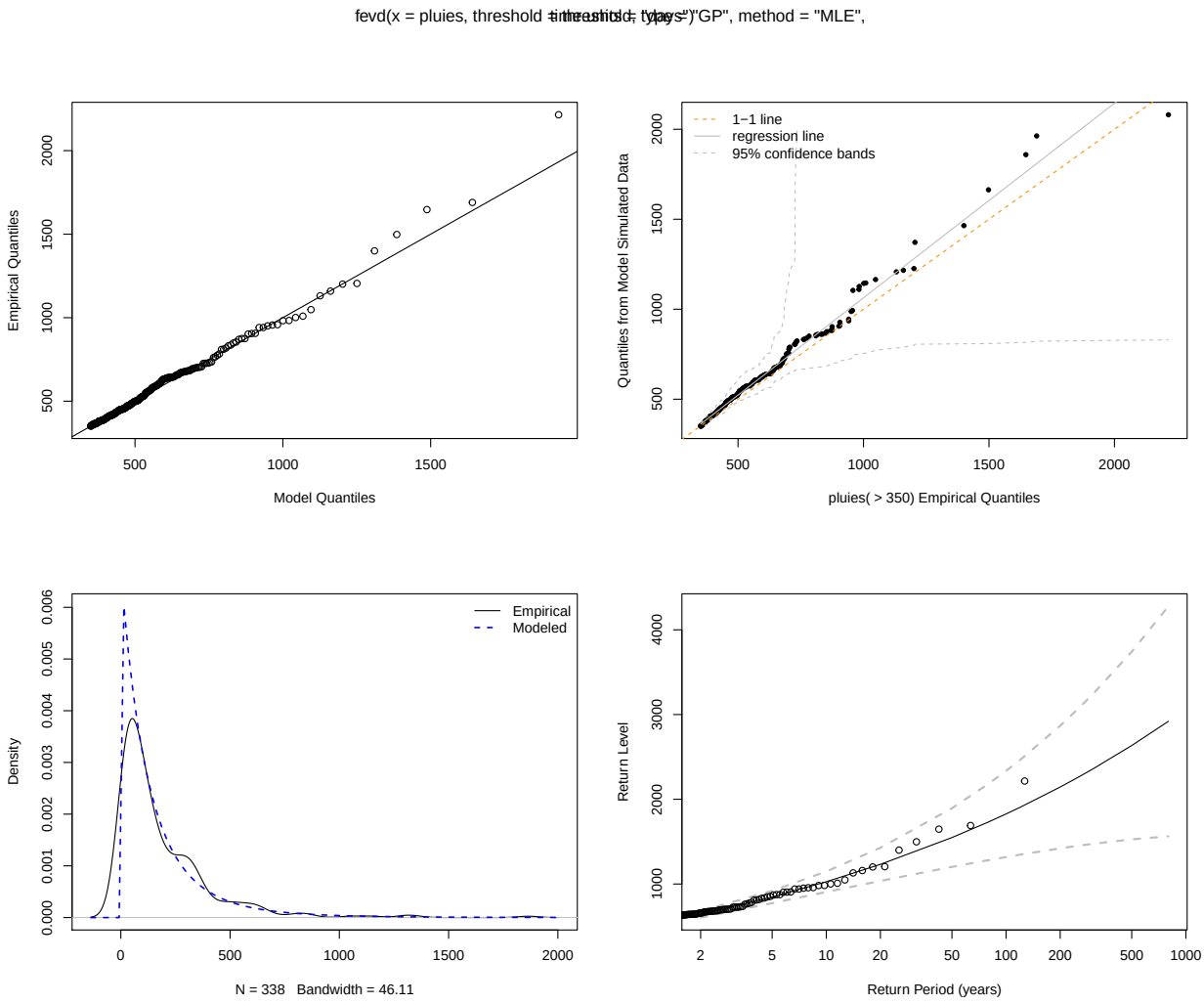
Avec la méthode du maximum de vraisemblance, nous avons un γ positif, égal à 0,19. Les maxima auront tendance à accroître de façon convexe sans point terminal.

```
##
## fevd(x = pluies, threshold = threshold, type = "GP", method = "MLE",
##     time.units = "days")
##
## [1] "Estimation Method used: MLE"
##
##
##  Negative Log-Likelihood Value:  2093.364
##
##
##  Estimated parameters:
##       scale       shape
## 148.5820618   0.1898363
##
##  Standard Error Estimates:
##       scale       shape
## 12.75821863  0.06750218
##
##  Estimated parameter covariance matrix.
##             scale        shape
## scale 162.7721426 -0.576349453
## shape  -0.5763495  0.004556544
##
```

```
##  AIC = 4190.727
##
##  BIC = 4198.373
```

Nous pouvons vérifier que la qualité d'ajustement est bonne. Les données s'alignent bien sur les graphiques quantile-quantile. La sous-estimation est très minime par rapport aux modèles précédants. La densité s'ajuste très bien aux données observées. Les données s'ajustent très bien sur la droite de return level plot.

Les résultats avec la lois d'excès sont très satisfaisants.



Le modèle prédit un niveau de retour plus élevé, avec 1.826 pour la période de retour de 100 ans et 3.065 pour la période de retour de 1000 ans. Cette prédiction semble beaucoup plus pertinente.

```
## fevd(x = pluies, threshold = threshold, type = "GP", method = "MLE",
##     time.units = "days")
##
## [1] "Normal Approx."
##
##                         95% lower CI Estimate 95% upper CI
## 100-year return level       1318.194 1826.845     2335.496
## 150-year return level       1380.337 2007.634     2634.931
## 1000-year return level      1574.361 3065.594     4556.826
```

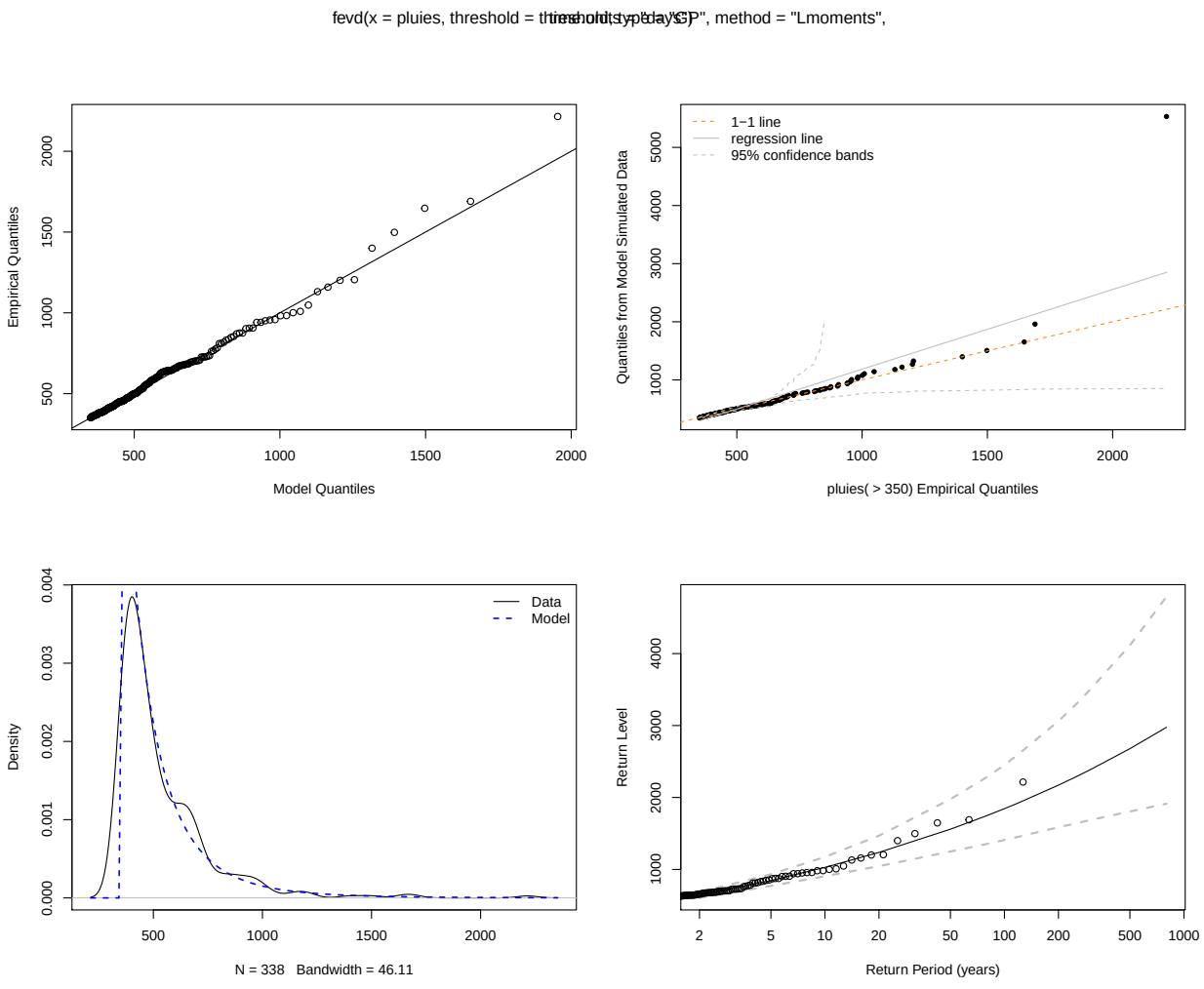
L'intervalle de confiance pour le paramètre γ est toujours positif. Nous n'avons pas d'incertitude du modèle.

```
## fevd(x = pluies, threshold = threshold, type = "GP", method = "MLE",
##     time.units = "days")
##
## [1] "Normal Approx."
##
##       95% lower CI    Estimate 95% upper CI
## scale 123.57641281 148.5820618  173.5877109
## shape   0.05753452   0.1898363    0.3221382
```

Estimation par la méthode des moments

Quant à la méthode des moments, l'estimation nous donne un γ positif, égal à 0,2. Il est très proche de γ estimé par la méthode du maximum de vraisemblance.

```
## fevd(x = pluies, threshold = threshold, type = "GP", method = "Lmoments",
##     time.units = "days")
## [1] "GP  Fitted to  pluies  using L-moments estimation."
##       scale       shape
## 147.3024342   0.1961344
```



Nous observons une bonne qualité d'ajustement. Les données sont très bien alignées sur les graphiques quantile-quantile. La densité est très bien ajustée. Nous avons tous les données qui sont presque sur la droite du return level plot. L'intervalle de confiance devient plus large que dans le cas de méthode avec le maximum de vraisemblace puisque γ est légèrement plus grand d'où l'accroissement plus rapide avec la période de retour.

```
## fevd(x = pluies, threshold = threshold, type = "GP", method = "Lmoments",
##     time.units = "days")
##
## [1] "Parametric Bootstrap"
## 502  iterations
##
##               2.5% Estimate    97.5%
## 100-year  1461.438 1844.728 2476.316
## 150-year  1564.013 2030.616 2828.178
## 1000-year 2092.904 3126.714 5258.426
```

Le modèle donne un niveau de retour similaire à celui de maximum de vraisemblance. Il est égal à 1.844 pour la période de retour de 100 ans et 3.126 pour 1000 ans. Légèrement plus grand car l'augmentation de maxima est plus rapide avec un γ plus grand.

4. Conclusion

Entre la loi d'extremum généralisée et la loi de pareto généralisée, la dernière extrapole mieux les valeurs extrêmes. Comme l'échantillon utilisé est plus riche en nombre d'observations et en qualité d'information, la loi de pareto généralisée a donné un résultat plus pertinent avec moins d'incertitude par rapport à la loi d'extremum généralisée.

Pour la période de retour de 100 ans, la loi d'excès a donné un niveau de retour plus proche à la valeur maximale sur les 127 ans que la loi de maxima.

Pour la période de retour de 1000 ans, la loi d'excès a donné un niveau de retour beaucoup plus élevé que la valeur maximale sur les 127 ans par rappoart à la loi de maxima. En effet, la loi de maxima donne un niveau de retour à l'entour de 2.400 et 2.500, proche à la valeur maximale 2.215 que nous observons pendant les 127 ans. La période de retour étant presque 8 fois plus grande que la période observée, nous espérons voir une valeur qui est beaucoup plus grande. Dans ce contexte, la loi d'excès qui nous a donné un niveau de retour supérieur à 3.000 semble plus pertinente.

De plus, la loi d'extremum généralisée présente une incertitude de modèle entre *Fréchet* et *Gumbel* et la loi de *Gumbel* estime des valeurs encore moins pertinent où le niveau de retour pour la période de retour de 1000 ans est inférieure à la valeur maximale observée sur les 127 ans.

Nos valeurs extrêmes semblent donc bien être dans le domaine d'attraction de *Fréchet* où il croit de façon convexe. La loi de pareto généralisée estime un γ plus grand qui fait que son niveau de retour accroît plus vite.

```
##                 100 ans  1000 ans MAX observé
## GEV - MLE     1607.505  2416.156        2215
## GUMBEL - MLE  1412.624  1885.199        2215
## GEV - Moments 1651.503  2562.380        2215
## GP - MLE      1826.845  3065.594        2215
## GP - Moments  1844.728  3126.714        2215
```

Nous pouvons également confirmer la pertinence de la loi de pareto généralisée avec le return level plot. Les données sont très bien ajustées avec la loi de pareto généralisée. Toutes les données se trouvent dans l'intervalle de confiance et quelques observations qui ne sont pas sur la droite restent toujours très proche de la droite.

Avec la loi d'extremum généralisée, les données s'ajustent bien dans la globalité mais l'incertitude devient de plus en plus grande avec la période de retour puisque les données sortent de l'intervalle de confiance à droite.

Pour notre jeu de données, la loi d'excès semble bien adapté pour estimer les valeurs extrêmes. Pour la loi, la méthode de maximum de vraisemblance bien que celle des moments s'ajustent très bien.