Verificatie van Ensembleverwachtingen

Leerzame experimenten met een eenvoudig niet-lineair model

Seijo Kruizinga en Kees Kok (KNMI)

Ensemble verwachtingen worden in eerste benadering gekenmerkt door twee grootheden: het gemiddelde van het ensemble en de spreiding van het ensemble. In het vorige artikel (Meteorologica, dec. 2003) hebben we gezien dat de spreiding (standaarddeviatie) van het ensemble hetzelfde groeipatroon en dezelfde variatie van dag tot dag vertoont als de fout(sd) (waarmee we dus aanduiden de standaarddeviatie van de fout) van de controle verwachtingen (onverstoorde verwachtingen startend vanuit de waargenomen begintoestand). Het ligt dus voor de hand om de spreiding van het ensemble te gaan gebruiken als een maat voor de skill van de verwachting. Echter de overeenkomst in de hiervoor genoemde twee eigenschappen is echter niet voldoende om van een succesvolle skill-verwachting te kunnen spreken. Daarvoor is ook nodig dat het ensemble breed is op momenten dat de control verwachting een lage kwaliteit heeft en juist smal bij een hoge kwaliteit van de verwachtingen. Er zijn meerdere mogelijkheden om deze vereisten te toetsen. De methode die we in dit artikel willen gaan gebruiken is het zogenoemde Rank Histogram, ook wel Talagrand diagram genoemd.

Het Rank Histogram hangt samen met de welbekende pluim van de ensemble verwachtingen. Als we in de pluim van bijvoorbeeld temperatuurverwachtingen bij een gegeven verwachtingstijd, bijvoorbeeld dag vijf, een verticale lijn trekken dan komen we van onder af eerst het ensemble met de laagste temperatuur tegen, daarna de volgende, enzovoorts tot de hoogste temperatuur. Op deze verticale lijn kunnen we achteraf ook de opgetreden temperatuur plotten en wil het ensemble concept nu enige betekenis hebben dan verwachten we dat soms de opgetreden temperatuur bij de onderste ensemble waarde ligt, een andere keer bij één van de middelste waarden en ook af en toe bij de hoogste waarde. In hoeverre hieraan wordt voldaan toetsen we feitelijk met het Rank Histogram.

In dit artikel willen we:

- het statistische concept van het Rank Histogram introduceren
- aan de hand van Rank Histograms aantonen dat ensemble verwachtingen in staat zijn om de skill van een verwachting vooraf aan te geven
- de algemene karakteristieken van het Rank Histogram als functie van de verwachtingstijd onderzoeken

Uiteraard maken we weer gebruik van onze referentieatmosfeer en het perfecte model. Van de referentieatmosfeer beschikken we echter uitsluitend over waarnemingen met een fout(sd)=0,01 en er zit geen bias in onze waarnemingen.

Verificatie met behulp van het Rank Histogram

Evenals in het vorige artikel zullen we onze analyse baseren op de gegevens, zowel verwachtingen van het ensemble als waargenomen waarden, die we voor X_1 beschikbaar hebben. We roepen in herinnering dat we daar het geordende ensemble hebben geïntroduceerd bestaande uit de verwachtingen voor X_1 van alle 50 Ensemble leden maar dan geordend naar grootte met de kleinste als eerste. Deze 50 geordende waarden delen de range van waarden die X_1 aan kan nemen op in 51 niet overlappende intervallen. Het eerste interval loopt van min-oneindig tot de kleinste waarde. Het tweede interval loopt van groter dan of gelijk aan de kleinste waarde tot de daaropvolgende enzovoorts tot het 51^e interval dat alle mogelijke uitkomsten omvat groter dan of gelijk aan de hoogste waarde uit

het ensemble. Deze intervallen, genummerd van 0 tot 50, worden ook wel bins genoemd. De basis-veronderstelling van de ensemble techniek is nu dat deze 50 ensemble waarden een willekeurige trekking zijn uit de conditionele kansverdeling van de werkelijke waarde van X_1. Als deze veronderstelling juist is dan kan op grond van statistiek worden aangetoond dat de werkelijke waarde van X_1, gemiddeld genomen, even vaak in ieder van de bins wordt aangetroffen. Uiteraard zal bij één ensemble run (bestaande uit 50 leden) voor een gegeven dag de opgetreden waarde slechts in één van de bins worden gevonden. We kunnen de veronderstelling dus uitsluitend toetsen op basis van meerdere runs en in het algemeen zijn daarvoor heel veel runs nodig.

Toetsing van de ensembles voor dag 15

Uiteraard is een skill-verwachting vooral interessant in het gebied van de foutengroeicurve waarin veel dag tot dag variatie optreedt in de fout(sd) in de verwachting. Voor de experimenten met een beginfout(sd)=0,01 is dat omstreeks dag 15 in de verwachting. Voor ieder van de 10000 waargenomen toestanden van de referentieatmosfeer hebben we een ensemble verwachting gemaakt uitgaande van de waargenomen toestand van 15 dagen eerder. Hieruit hebben we voor iedere dag de geordende ensemble waarden van X_1 genomen en vervolgens genoteerd welke bin de waargenomen waarde bevat. Daarna hebben we geteld hoe vaak iedere bin voorkomt in de set van 10000 dagen. Als de veronderstelling omtrent het ensemble correct is dan verwachten we in iedere bin 10000/51 keer de verifiërende waarneming (in procenten 100/51=1,96%). In figuur 1a vinden we het resultaat voor de dag 15 verwachting. We zien in deze figuur dat de frequenties van de bins willekeurig spreiden rond 1,96%. Op grond van statistische berekeningen mogen we ook verwachten dat deze percentages spreiden van omstreeks 1,76 tot 2,24%. Deze figuur bevestigt dus dat gemiddeld genomen de spreiding van de ensemble waarden een goede maat is voor de spreiding van de verifiërende waarnemingen rond de verwachting. In statistische termen zeggen we dan dat er op grond van deze figuur geen aanleiding is om de basisveronderstelling van de ensemble verwachting te verwerpen, of anders gezegd, de spreiding van het ensemble geeft een goed beeld van de spreiding die er op kan treden in de waargenomen waarde behorend bij het ensemble. Echter dit geldt alleen nog maar voor het geval dat we een frequentieverdeling opmaken over alle gevallen. Het zou best eens kunnen zijn dat bij smalle ensembles er veel opgetreden waarden in de buitenste bins voorkomen en bij brede ensembles juist veel opgetreden waarden in de centrale bins. Om na te gaan of dit al of niet het geval is hebben we ook een Rank Histogram gemaakt voor de 2000 smalste ensembles (ensembles met een kleine spreiding) en eveneens voor de 2000 breedste ensembles. In de figuren 1b en 1c vinden we de bijhorende frequentieverdelingen. Nog steeds zijn de frequenties voor alle bins ongeveer even groot, omstreeks 1,96%. De spreiding in de frequenties van bin tot bin is wel groter maar dat hangt samen met het kleinere totaal aantal, 2000 in plaats van 10000. Voor de situatie op dag 15 kunnen we dus concluderen dat de verifiërende waarneming van X_1 zowel bij smalle als bij brede ensembles in het gebied dat door het ensemble wordt aangegeven wordt gevonden, oftewel, bij dag 15 geven de ensembles correcte informatie met betrekking tot de potentiële skill van de verwachting. De splitsing naar smalle en brede ensembles is een strengere toetsing dan gebruikelijk in de literatuur voorkomt. Vaak is dat echter in praktijk situaties ook niet mogelijk wegens gebrek aan voldoende data. Dat een strengere toetsing wel zin heeft zien we aan de volgende voorbeelden.

Toetsing van de Ensembles voor dag 30 en dag 3

In de figuren 2a, 2b en 2c vinden we overeenkomstige Rank Histogrammen, zoals hiervoor beschreven bij dag 15, voor de dag 30 verwachting. Wederom ziet het totale Rank Histogram er goed uit. Echter bij de smalle en brede ensembles zijn er duidelijke afwijkingen. Bij de brede ensembles wordt de verifiërende waarneming meer in het centrum van het ensemble gevonden, bij smalle ensembles wordt de verifiërende waarneming juist meer aan de randen gevonden. Dit betekent dat de samenhang met de potentiële skill zwakker wordt. Hoeveel zwakker is op grond van deze plaatjes moeilijk te kwantificeren. Het is niet bekend of dit effect bij atmosferische verwachtingen ook optreedt maar dit experiment geeft duidelijk aan dat het niet voldoende is om uitsluitend naar het totale Rank Histogram te kijken.

De overeenkomstige Rank Histogrammen voor dag 3 vinden we in de figuren 3a, 3b en 3c. In het totale Rank Histogram zien we hier duidelijk een concentratie van de verifiërende waarneming aan de randen. Bij de smalle ensembles is dat nog veel sterker maar bij de brede ensembles is dat effect verdwenen. De oorzaak is nu echter niet gelegen in het falen van de ensemble techniek maar in de wijze waarop we het Rank Histogram samenstellen. Volgens de basisveronderstelling geeft het ensemble aan waar we de werkelijke (zonder waarnemingsfout) verifiërende waarde van X_1 zullen vinden. Bij het samenstellen van het Rank Histogram zijn we echter uitgegaan van de verifiërende waarneming en daar zit een extra spreiding in als gevolg van de waarnemingsfout. De standaard-deviatie van de waarnemingsfout is zoals eerder gezegd 0,01 en de standaarddeviatie van het ensemble is bij dag 3 omstreeks 0,025. Deze twee standaarddeviaties zijn van vergelijkbare grootte vandaar dat de verifiërende waarneming vaker buiten het ensemble interval wordt gevonden. Dit verklaart ook waarom het effect bij smalle ensembles zoveel sterker is. Bij dag 15 en dag 30 is de standaarddeviatie van het ensemble zeker 100 maal groter dan de standaarddeviatie van de fout in de waarneming en daar is dus het effect verwaarloosbaar. Zie hiervoor ook Saetra et.al. (2002).

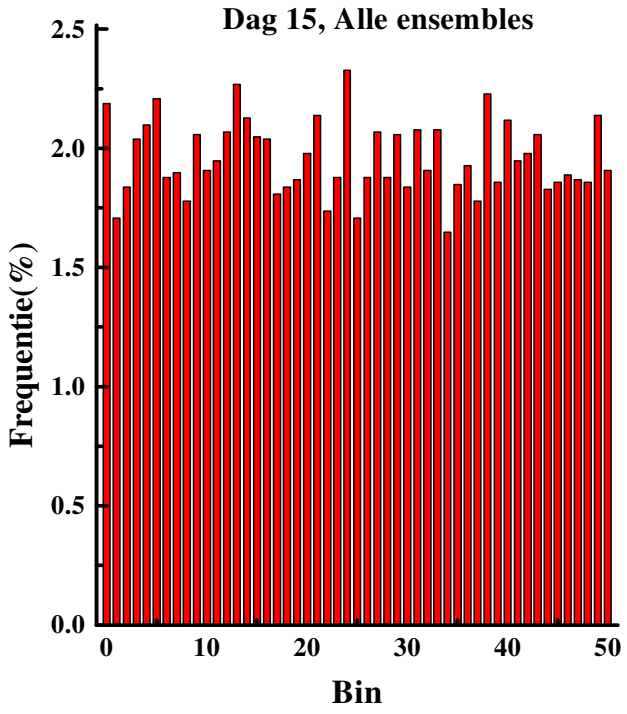
In de voorgaande analyses hebben we gebruik gemaakt van runs uitgaand van een initiële fout met een standaarddeviatie van 0,01. In de atmosferische praktijk is de initiële fout relatief veel groter en daar verwachten we overeenkomstige effecten ook bij andere verwachtingstijden. Het verzadigings-effect bijvoorbeeld al bij dag 10, het midden van de groeicurve omstreeks dag zeven en het begineffect bij dag twee of drie.

Slot

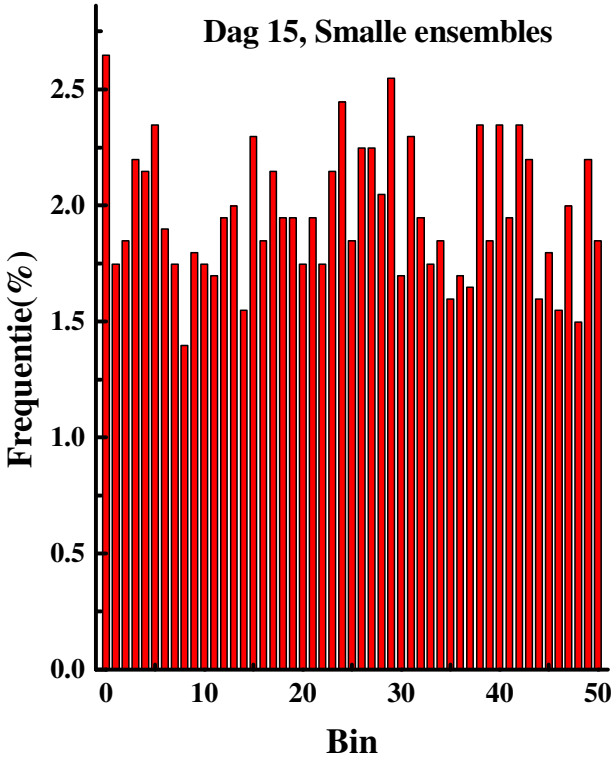
In het voorgaande hebben we, aan de hand van het Rank Histogram, laten zien dat de ensemble techniek in staat is om op voorhand een uitspraak te doen over de kwaliteit van de verwachting. Van zo'n Rank Histogram verwachten we dat het "vlak is" oftewel dat alle bins even vaak voorkomen. Echter, we hebben duidelijk gezien dat dit voor de korte termijn, zolang de verwachtingsfout van dezelfde orde van grootte is als de waarneemfout, niet opgaat en ook niet kan opgaan. Bij de langere termijn wordt hier wel aan voldaan. Echter meer gedetailleerd onderzoek geeft aan dat in het verzadigingsgedeelte van de foutengroei de relatie tussen ensemblespreiding en skill zwakker wordt. Deze algemene karakteristieken zijn van belang bij de beoordeling van Rank Histogrammen uit de praktijk. Daarnaast zou het nuttig zijn om te weten hoe het Rank Histogram reageert op fouten in het verwachtingssysteem. Deze laatste aspecten kwamen in dit verhaal niet aan de orde. In het artikel van Saetre et. al. (2002) wordt hier wel aandacht aan besteed.

Literatuur

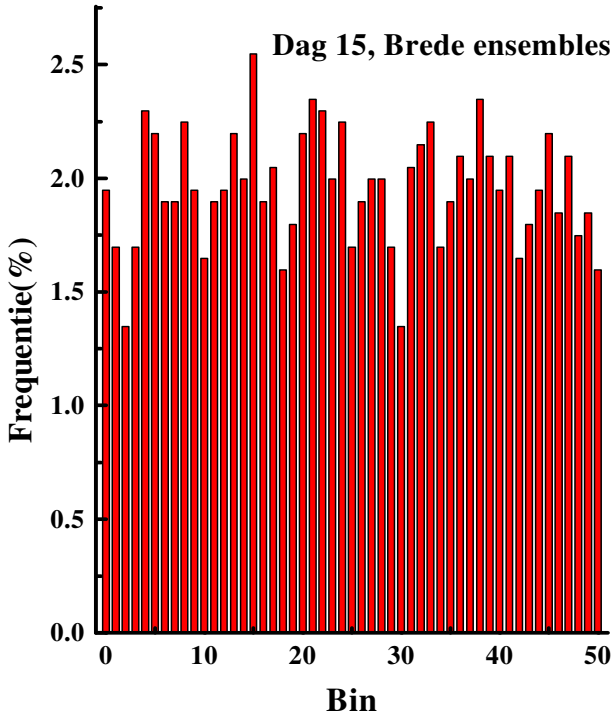
Saetra, O., J-R. Bidlot, H. Hersbach and D. Richardson, 2002: Effects of observation errors on the statistics for ensemble spread and reliability, ECMWF Research Department Tech.Mem. 393, 12p.



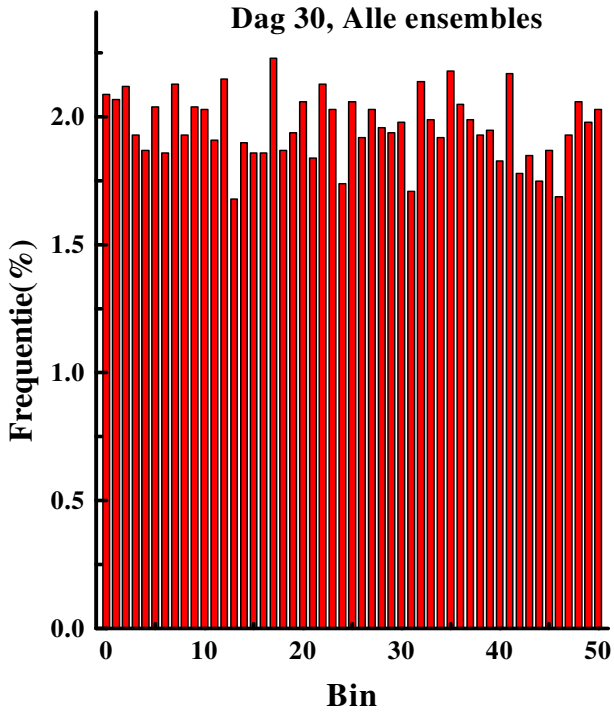
Figuur 1a Rank Histogram van alle Ensembles voor dag 15 met een beginfout(sd)=0,01.



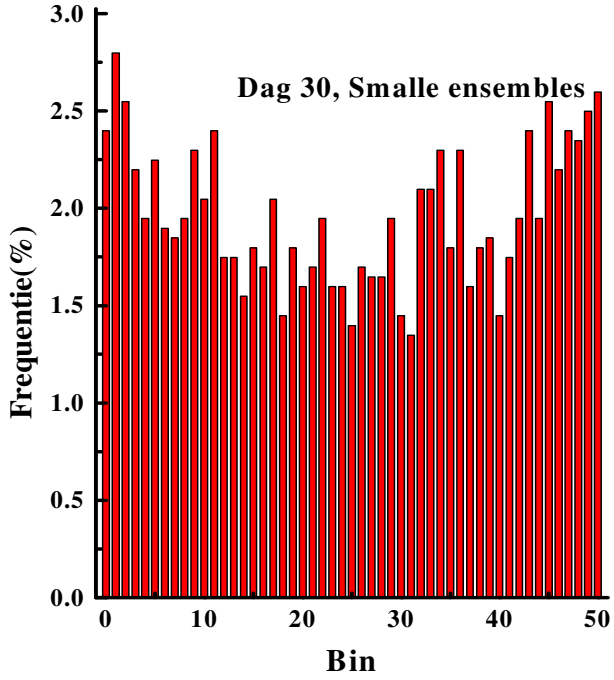
Figuur 1b Rank Histogram van de 2000 smalste Ensembles voor dag 15 met een beginfout(sd)=0,01



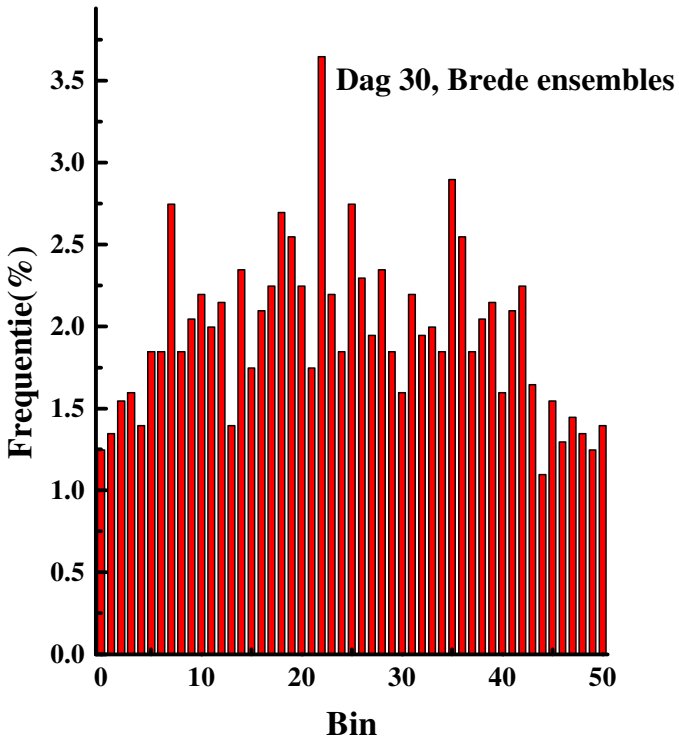
Figuur 1c Rank Histogram van de 2000 breedste Ensembles voor dag 15 met een beginfout(sd)=0,01



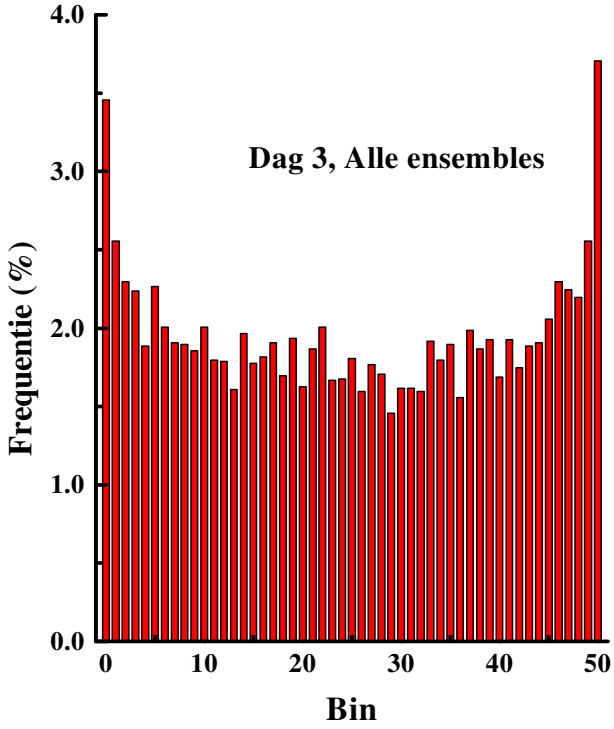
Figuur 2a Rank Histogram van alle Ensembles voor dag 30 met een beginfout(sd)=0,01



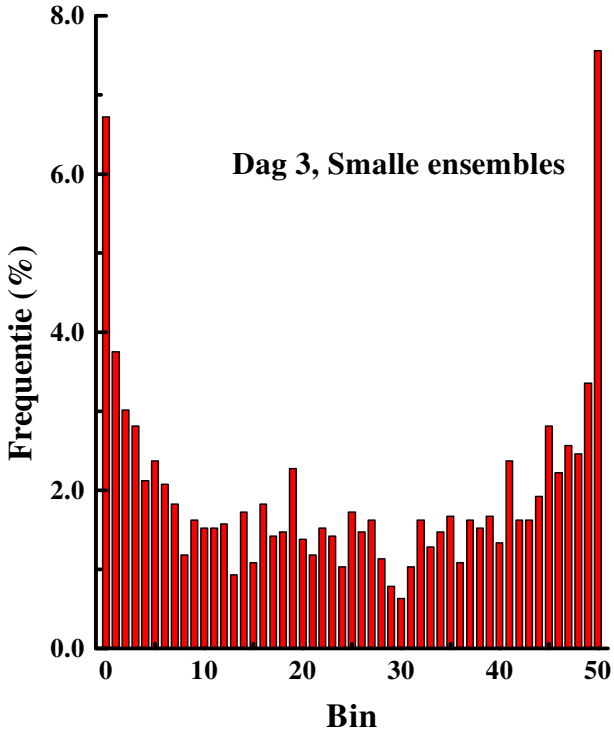
Figuur 2b Rank Histogram van de 2000 smalste Ensembles voor dag 30 met een beginfout(sd)=0,01.



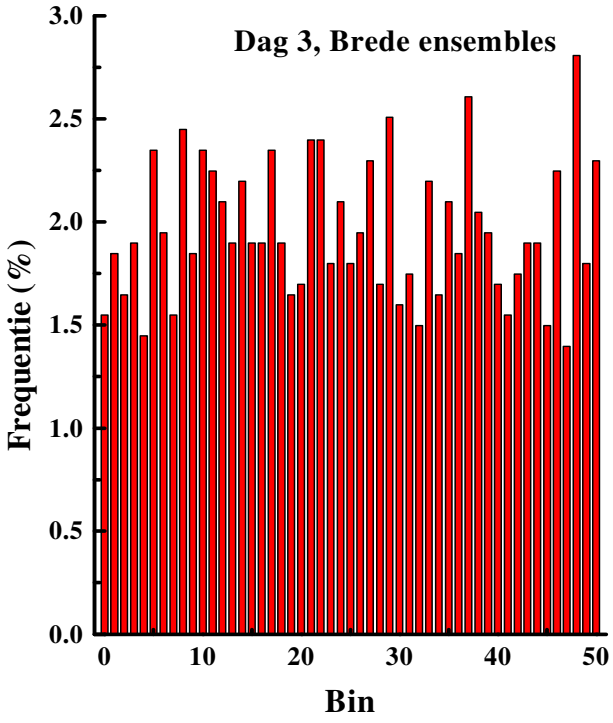
Figuur 2c Rank Histogram van de 2000 breedste Ensembles voor dag 30 met een beginfout(sd)=0,01.



Figuur 3a Rank Histogram van alle Ensembles voor dag 3 met een beginfout(sd)=0,01.



Figuur 3b Rank Histogram van de 2000 smalste Ensembles voor dag 3 met een beginfout(sd)=0,01.



Figuur 3c Rank Histogram van de 2000 breedste Ensembles voor dag 3 met een beginfout(sd)=0,01.