Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 July 2021 doi:10.20944/preprints202107.0666.v1

Article

A Re-Evaluation of the Swiss Hail Suppression Experiment using Permutation Techniques shows Enhancement of Hail Energies when Seeding

Armin Auf der Maur [1] and Urs Germann [2,*]

[1] Schachenstrasse 18, 6030 Ebikon, Switzerland; arminaufdermaur@yahoo.de
[2] MeteoSwiss, Locarno-Monti, Switzerland; UrsGermann@meteoswiss.ch
* Correspondence: UrsGermann@meteoswiss.ch

Received: date; Accepted: date; Published: date

Abstract: Grossversuch IV is a large and well documented experiment on hail suppression by silver iodide seeding. The original 1986 evaluation remained vague, although indicating a tendency to increase hail when seeding. The strategy to deal with distributions of hail energy far from normal was not optimal. The present re-evaluation sticks to the question asked and avoids both misleading transformations and unsatisfactory meteorological predictors. The raw data show an increase by about a factor of 3 for the hail energy when seeding. This is the opposite of what seeding is supposed to do. The probability to obtain such a result by chance is below 1%, calculated by permutation and bootstrap techniques applied on the raw data. Confidence intervals were approximated by bootstrapping as well as by a new method called "correlation imposed permutation" (CIP).

Keywords: hail prevention, non-normal distributions, permutation, bootstrap, confidence intervals

1. Introduction

Damage on crops and other objects by hail is a disaster which lead especially farmers to protect themselves by various means. A comprehensive review on hail suppression by different methods was done by Wieringa and Hollemann [1]. A most recent study by Rivera et al. [2] confirmed the uncertainty about a possible benefit of 60 years of hail suppression in Mendoza (Argentina) and points to several rather recent studies in other countries. Still nowadays silver iodide seeding from airplanes is a commercially available practice. A convincing proof of the benefit of such actions is not known to the present authors. Contrarily, the present study shows evidence that one of the best documented experiments, Grossversuch IV described by Federer [3] had an opposite result. The effect is large, about a factor of 3 and hardly explained by chance. To our knowledge it is for the first time that a statistically sound indication shall be given for the increase of hail energy when seeding thunderstorms by silver iodide. This is an important result which should induce enterprises to reconsider their efforts to suppress hail by seeding.

Grossversuch IV was conceived to verify the benefit of some operational programs of hail suppression in the former Russia and other countries in East Europe, see [3] p. 917. The idea is to increase artificially the number of hail embryos in order to reduce the size of hailstones by the competition for the available supercooled water, according to [4]. Unfortunately [3] failed to give a concise answer, it concluded "...a majority of the evaluations suggest some trend to larger seeded hail energy and larger seeded-hail area values...". A main problem was the distribution of the response variable hail energy, which was far from normal. Different ways were followed up in [3] to cope with the problem, but the confirmatory test announced in advance was not satisfactory for several reasons:

1. The magnitude of sc, the treatment variable called seeding coverage $0 \leq sc \leq 1$, contains the information how well seeding was done. But this information was not used. Instead, sc was replaced by what was planned, although some 20% of the cells planned for seeding were not at all seeded.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 July 2021 doi:10.20944/preprints202107.0666.v1

2. The response variable hail energy on the ground was converted to its logarithm. It will be shown how misleading this can be.
3. Some evaluations used a predictor based on meteorological or other data. This reduced the importance of severe hailstorms and introduced errors in the statistical analysis.
4. Some evaluations based on data from the hailpads are shown to be not representative enough to calculate hail energies.

The 1986 study [3] contained also an exploratory evaluation with neither predictor nor logarithmic transformation applied. It revealed a ratio of 2.2 for the hail energy of seeded over non-seeded cells. Pure chance had only a probability of 2% to produce this result according to the underlying $C(\alpha)$ test. The authors attributed this result to the multiplicity effect "... which means that some out of a number of tests turn out significant by pure chance...". The review on hail suppression experiments by [1] did not mention this result. But the present re-evaluation confirms even more stringently that this exploratory analysis was on the right track.

The evaluation of an experiment should be defined before the results are known in order to prevent searching for some accidentally significant results. This is an important point with respect to the present re-valuation. Our answer is that we stick as closely as possible to the original question about hail suppression or enhancement by silver iodide seeding. Although questions and answers can be slightly different a homogeneous picture will emerge, different ways of evaluation lead to similar answers. All the results show an increase of hail energy when seeding. The probabilities $P(H_0)$ of an accidental result are about 0.4% if all the available information is used. The permutation method applied to calculate $P(H_0)$ copes well with non-normal distributions. It was also used and explained by Federer [3], p. 925.

The calculation of confidence intervals CI for non-Gaussian data is still a challenge [5]. The current bootstrapping method introduced by Efron [6] resamples the data outcome and treatment in pairs, leaving some out and selecting others twice or more times. A new method is presented here using all the data just once, permuting the associations in a way to impose the original correlation. Fortunately both methods agree for the hail data so that there is no need to decide which model simulates better the experiment. Permutation and bootstrap diverge for certain data. This issue is treated in the appendix.

The seeding effect is large, although not precise. The ratio $rr = 3$ of seeded over non-seeded could be only 2 within one standard deviation (std).

The increase of hail energy when seeding is remarkable and opposite to what was expected. The hypothesis, that many ice particles created by seeding compete for the available super-cooled water and suppress the formation of large hail, may work in small thunderstorms, but in large systems with plenty of supercooled water seeding may rather open ways to additional hail formation. Some ideas in this direction are presented.

A quantitative model of the seeding effect is beyond the scope and extent of this investigation. It concentrates rather on the challenging statistical issues which were treated in the study of 1986 [3] ambiguously. The interesting design, the extent of the expensive experiment and the quality of the available data call for a statistical re-evaluation.

2. The hail suppression experiment Grossversuch IV

Grossversuch IV is a randomized experiment performed in the years 1977–1982. The goal was to find out whether seeding thunderstorms by silver iodide according to a Russian procedure using Oblako rockets would change hail energy at ground in a statistically significant way.

The experimental region covering about 1300 km^2 was surveyed by radar and by hailpads. On 83 experimental days 253 convective cells were found to comply with the conditions for seeding and the hail energy on ground E_{GR} was estimated by radar. The treatment seeding or not seeding was decided according to a randomized daily scheme. A visualization of the data is shown in Figure 1. The hail energies E_{GR} are stratified by the lifetime of the cells, i. e. the time between the criterion of seeding

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 July 2021 doi:10.20944/preprints202107.0666.v1

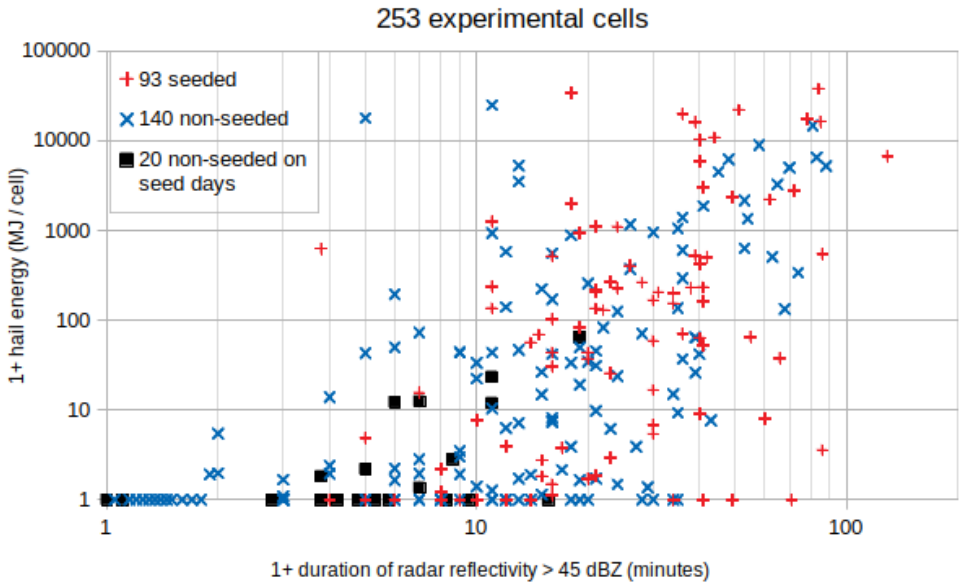

Figure 1. Visualization of the data from the Swiss hail suppression experiment "Grossversuch IV" [3] that builds the basis of the re-evaluation presented in this paper. For better readability some overlapping points have been slightly separated on the x-axis.

first and last met. The lifetime of the cells is typically 10 to 100 minutes. Some of the shorter lifetimes may be due to cells moving into or out of the experimental zone.

Sulakvelidze [4] described the concept and procedure of seeding. Rockets containing silver iodide are shot into convective cells as soon as the radar reflectivity exceeds 45 dBZ. An Oblako rocket is aimed every five minutes at about the $-5C$ isotherm into the center of the cell as long as the criterion >45 dBZ is sustained. In asymmetric systems the targets could be feeder clouds or the forward overhang. Sometimes smaller rockets of the type PGIM were also used, four PGIM instead of one Oblako. The seeding technique is based on a Soviet concept of creating a surplus of frozen particles competing for the available supercooled water. The expectation was that the additional ice embryos may deplete the supercooled water of the cloud, reducing therefore the size of the hailstones, see [3, p.918], [4]. The hypothesis involves also an "accumulation zone" of large supercooled drops (big drop zone). The existence and role of such zones in Grossversuch IV was not clarified.

Seeding was not at all perfect for several reasons [3, p.942]. In the six years 1977–1982 a total of 113 cells should have been seeded, of which 20 were not at all seeded and 22 did not reach $sc = 1/3$, the threshold specified for satisfactory seeding, see Figure 2. These 42 cells should have been excluded from evaluation according to the original design of the experiment [3, p.943]. At the time of evaluation it was decided to leave these 42 cells within the seeded group in order to avoid a bias towards an increased average when the number of seeded cells would drop from 113 to 71. The mistake was to give these many cases the full weight of perfectly seeded. The seeding quality is expressed by $0 < sc \leq 1$ according to the quotient of the number of Oblako rockets fired to those required. This is better interpreted as the fraction of the lifetime of a cell during which seeding was performed. It may be mentioned that the strong positive correlation between sc and E_{GR} was obvious [3, Figure 14], but this track was not followed up.

The main response variable in the 1986 study was the kinetic hail energy E_{GR} for each experimental cell, either derived from radar or measured on ground by two hailpad networks run by an Italian and a French group. The radar based data are preferred in the present study for several reasons. They are available for the whole period 1977–1982 and the radar may follow a seeded cell moving out of the hailpad networks [3, p. 946]. Furthermore it will be shown that the scarce sampling of 0.1 m^2 per

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 July 2021 doi:10.20944/preprints202107.0666.v1

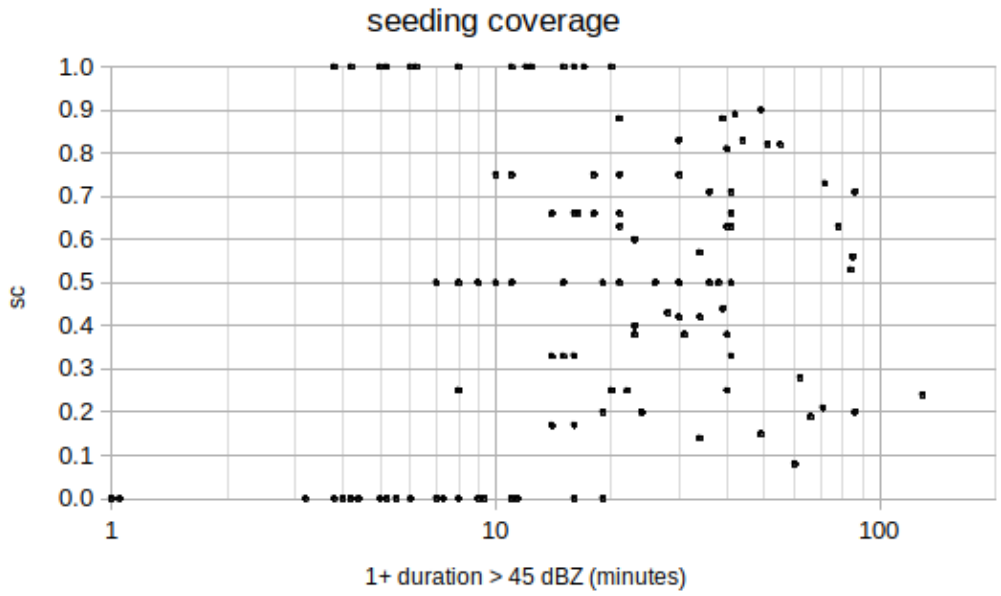

Figure 2. Visualization of the seeding coverage versus the duration of cell lifetime, which is defined as the time with radar reflectivity exceeding 45 dBZ. The dots correspond to the 113 cells on the days that have been selected for seeding in the randomization process.

hailpad representing 3.8 to 4 km^2 and maybe other errors led to stochastic variations which made it improbable to reach statistical significance for the demanding variable E_{GR}.

The radar used to calculate E_{GR} had a wavelength of 10.1 cm and was equipped with an antenna of 4.3 m diameter making a full rotation every 6 seconds. The day-to-day calibration was made with a microwave generator and absolute calibration was achieved by comparison with data from rain distrometers and hail spectrometers. To obtain an estimate of the uncertainty of E_{GR}, Waldvogel [7] used data collected by a hail spectrometer. Total energies obtained by converting the measured spectra into radar reflectivity Z and then into energies using a general $Z - E_{GR}$ relation agree with energies obtained directly from the spectra to better than 25%. For a detailed presentation of the measurements of Grossversuch IV and studies of data quality and error sources see [7–9] and [10].

The 1986 study based the confirmatory test on the logarithmic transformation $\ln(E_{GR}+1)$ because of the extremely high skewness of the distribution [3, p. 920–921]. No doubt hail energy on ground E_{GR} is a well chosen physical parameter to represent potential damage independent of the season and type of crops. This link is sacrificed by a logarithmic transformation. It removes the importance of severe cases. Exactly these cases are the interesting ones for an effective hail suppression - or the opposite.

Besides the treatment variable sc, a predictor variable f was sometimes added in the form of $\ln(E_{GR}+1) - f$, corresponding to $E_{GR} \cdot \exp(-f)$. This was done in the hope to reduce stochastic variations. But this procedure removes also the weight of large hailstorms and it can change the results substantially. Different predictors were derived from preliminary data, from data of Grossversuch IV, from meteorological data or from values of a control area. One of these predictors f found its way into the appendix of [3]. This one is responsible for a fictitious decrease of E_{GR} when seeding because f happened to be correlated with sc, counteracting the correlation between $\ln(E_{GR}+1)$ and sc. A real correlation between a meteorological predictor f and sc would be worrying. Fortunately the correlation observed for the logarithmic version f vanishes in the dimension of hail energies $\exp(f)$. When using $\ln(E_{GR}+1)$ alone without f for 93 really seeded and 160 non-seeded cells, a positive correlation with a slope of 1.54 at a significance level of $P(R|H_0) < 0.01\%$ would have been obtained. Using both f and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 July 2021 doi:10.20944/preprints202107.0666.v1

113 planned seeded and 140 non-seeded cells turns the positive correlation to negative with a slope of -0.31 and an insignificant $P(R|H_0)$ = 23%.

We think that keeping the evaluation simple and transparent is better than trying to reach statistical significance by the introduction of a secondary predictor beside *sc* with all resulting complications, especially when this predictor is not at all reliable. The authors introducing such predictors admit that "predicting hailfall is still an unresolved task" [3, p. 945]. A precise predictor could tell more about the type of the seeding effect (constant or rather stochastic), but in this case it turned out to remove the weight of the most severe hailstorms.

Most interesting is the exploratory analysis with not transformed hail energies and without an additional predictor [3, p. 945–946, Table 21], although the analysis based on a $C(\alpha)$-test involves two not granted assumptions: the fit of the data by a Gamma distribution and a fixed multiplicative effect of seeding. The evaluation of four different measurements (three radar-based and one ground-based) reveals an "increased seed / no-seed ratio and $P(H_0)$ values that indicate a significant- or almost-significant effect". In other words, when using the raw data, an increase of the hail energy in the case of seeding is found. The ratio *rr* is 2.2 and $P(rr|H_0)$ = 1.9% for E_{GR}, comparable to our re-evaluation when, for comparison with [2], the 20 planned but not seeded cases were transferred to the seeded group (line 5 in Table 2).

Measurements of hail energies by an Italian and a French group running a network of 333 hailpads, each 0.1 m^2 large and with a mesh area of 3.8 to 4.0 km^2 are found in [3, appendix]. The results correlate with those from the radar observation but the stochastic variations are too large to reach statistical significance for hail energies. Evidence for this statement is given later. More reliable are the results for a less demanding variable such as the area touched by hail. An increase by a factor of 1.8 was reported for seeding [3, Table 13, S_G].

3. The present re-evaluation

3.1. The variables and parameters

The present study is based on data found in [3, appendix]: the hail energy on the ground $\ln(E_{GR}+1)$, reconverted to E_{GR}, the seeding coverage *sc*, the beginning t_0 and the end t_f of the seeding criterion met within the experimental area. The lifetime of a cell $t = t_f - t_0$ serves to stratify the data for figures or to convert *sc* from cells to days. As randomization was done for days, the data given for cells had to be converted to the values relevant for the 83 experimental days. For the hail energy it is the sum of E_{GR} for each day. More complicated is the variable *sc*. The total number of rockets fired is given in [3, appendix], but the scores include the four times less effective PGIM type of rocket, which is confusing. Therefore the present calculations are based on *sc* and *t* to calculate *sc* for each day by $\sum(sc_i \cdot t_i)/\sum t_i$. This is the really seeded fraction of the lifetime of all cells of a day.

We set the response variable $y = E_{GR}$ and the treatment variable $x = sc$. The sample size n is 253 cells or 83 days, whereas n_s is the number of seeded cells (93) or the number of days with at least one seeded cell (34). Our interest is in a couple of parameters which characterize the difference *dif* or the ratio *rr* of *y* between seeded and non-seeded cells or days. There is a direct access from the variables *y* and *x* to the parameters *dif* and *rr* by the average of the non-seeded cells or days *avn* and the weighted average of the seeded *avs*. Obviously the relation to the parameters is $dif = avs - avn$ and $rr = avs/avn$. The weighted seeded average *avs* is calculated in this way:

$$avs = \frac{\sum_{i=1}^{n}(y_i \cdot x_i)}{\sum_{i=1}^{n} x_i} \tag{1}$$

A practical, more or less self explaining code for such expressions is used in the free software "Octave", compatible with Matlab: $avs = sum(y.*x)/sum(x)$, where .* indicates a term by term multiplication, and $avn = sum(y(x=0))/length(x(x=0))$.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 July 2021 doi:10.20944/preprints202107.0666.v1

When later permutations are applied on x to calculate probabilities, a problem could arise for the parameter rr if $avn = 0$. This could happen for certain permutations when there are less non-seeded cases than cases with no hail. But this is not true for the hail data. Some hail is found within the non-seeded group for all permutations.

There is an elegant alternative to *avn* and *avs*: correlation and regression. A classical measure of association between E_{GR} and sc is the Pearson correlation coefficient R, a versatile parameter. Two means as well as 2 x 2 contingency tables can be interpreted as a special case of correlation. R is standardized as a product of two "studentized" variables resulting in $-1 \leq R \leq 1$.

$$R = \frac{1}{\sigma_y \cdot \sigma_x}(-\overline{y} \cdot \overline{x} + \frac{1}{n}\sum_{i=1}^{n}(y_i \cdot x_i)) \tag{2}$$

The sign of R is important. A negative sign points towards hail suppression and a positive sign towards increased hail energy when seeding. Correlation is the key to regression with a slope $R \cdot \sigma_y/\sigma_x$ and an intercept $\overline{y} - \overline{x} \cdot slope$, allowing to calculate alternative estimates of dif and rr. The difference dif is given in MJ per cell or per day, the ratio rr is dimensionless (in 2 x 2 tables known as risk ratio). The difference dif is just R multiplied by a constant:

$$dif = R \cdot \frac{\sigma_y \cdot \overline{x} \cdot n}{\sigma_x \cdot n_s} \tag{3}$$

The sample size n, the number of seeded cases $n_s = length(x(x > 0))$, the averages $\overline{y}$ and $\overline{x}$ as well as the std σ_y and σ_x do not change when x is permuted. Only the term $\sum_{i=1}^{n}(y_i \cdot x_i)$ is affected by permutation.

More delicate is the formula for $rr = 1 + (slope \cdot \sum_{i=1}^{n} x_i/n_s)/intercept$, because the intercept could become zero. This is explicitly shown in the following formula for rr:

$$rr = 1 + \frac{R \cdot n/n_s}{-R + R_{cr}} \tag{4}$$

The critical constant R_{cr} is

$$R_{cr} = \frac{\overline{y} \cdot \sigma_x}{\overline{x} \cdot \sigma_y} \tag{5}$$

R_{cr} of the hail data is 0.44 and 0.66 for cells and days, respectively. These values are not changed by permutations. A second critical point R_{c2} may be found at $rr = 0$, corresponding to $R_{c2} = -Rcr \cdot ns/(n - ns)$. When calculating probabilities for rr by permutations or bootstrap, R_i of every permutation i must be kept within these limits R_{cr} and R_{c2}. This does not change the medians of R and rr in the vicinity of $R = 0$ or $rr = 1$. Means, however, would be corrupted.

Table 1. Parameters *dif* in MJ per cell and risk ratio *rr* calculated by two models: regression or weighted average based on *avs*, *avn*. Conversion of *dif* for days to MJ/cell by the factor 83/253. The probabilities calculated later in section 3.2 are added.

model	n	$dif(MJ/cell)$	$P(dif\|H_0)$	rr	$P(rr\|H_0)$
regression	83	1610	0.38%	3.27	0.38%
regression	253	1583	0.38%	3.01	0.38%
avs, *avn*	83	1721	0.53%	3.01	0.87%
avs, *avn*	253	1942	0.29%	3.50	0.31%

Table 1 shows the agreement and differences when calculating dif and rr by regression or by weighted averages. Both take unsatisfactory seeding into account, but in a different manner. The weighted average *avs* neglects practically all of y when the corresponding x is close to zero. Regression is not affected by this kind of discontinuity. Therefore differences between the models must be expected. Ideally, rr should be equal for the 83 days and 253 cells, whereas dif is made to become comparable

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 July 2021 doi:10.20944/preprints202107.0666.v1

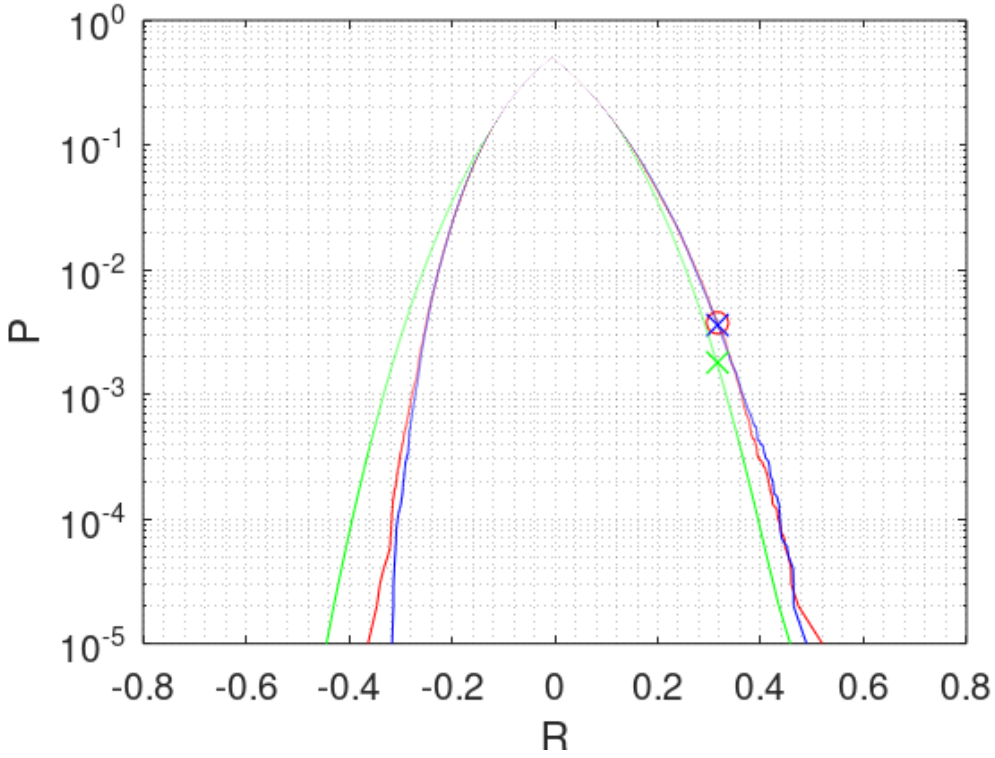

Figure 3. Probabilities that an observed correlation coefficient R is by pure chance, that is the null hypothesis H0 is true. Three curves show the cumulative distribution function of $min(P_i, 1-P_i)$ to read off $P(R|H_0)$ for the methods Fisher's z (green), permutation (blue) and bootstrap y (red). The underlying y are the hail energies of the 83 days, x the seeding coverage.

by converting *dif* per day to *dif* per cell by the factor 83/253. Hail cells are more interesting than days because the hail energy of cells can be compared to cells elsewhere, whereas for days such a comparison makes less sense. Table 1 reveals quite a difference between the models and an appreciably better agreement between days and cells for regression. Therefore the model regression is preferable.

It is important to note that the direct way by *avs* and *avn* is identical to regression when *sc* is simplified to a binary seeded, non-seeded. This advantage does not outweigh the loss of accuracy when discarding the detailed information contained in *sc*.

3.2. The calculation of probabilities

The crucial question concerns the probability $P(R|H_0)$. Could it be that the observed R, *dif* or *rr* would be due to chance? If this chance $P(R|H_0)$ is below the classical 2.5% in one of the two tails, the null hypothesis (H_0) is judged improbable. The task is to calculate the probability for the observed results assuming that H_0 is true.

Different methods will be compared with respect to the parameter R. One of the oldest is based on student's t or Fisher's z. The latter is simpler and a close approximation to the probabilities obtained by t.

$$z = 0.5 \cdot (n-3)^{-0.5} \cdot \ln\left((1+R)/(1-R)\right) \tag{6}$$

It should be noted that the original data $y = E_{GR}$ are not transformed, only R as part of the calculation of probability. If x and y are samples from normal distributions, z is a standard normal distribution. In this case $P(z)$ as well as $P(R)$ are known. The green line in Figure 3 shows the accumulated probabilities $min(P, 1-P)$ for the 83 hail days starting from both extremes of R. This way of plotting a

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 July 2021 doi:10.20944/preprints202107.0666.v1

cumulated distribution function (cdf) allows to use a logarithmic scale with adequate resolution and showing both tails, peaking at the median of R.

The green curve for P(z(R)) is symmetrical, which is not realistic for the hail data. As the sample E_{GR} is far from a normal distribution, combinative tests should be applied. The randomization test is such a test, characterized by the permutation of one variable. It was introduced by R. A. Fisher in 1924 according to [11, p. 3]. The confirmatory test of Grossversuch IV was a complicated version of the randomization test and regression in two dimensions [3]. It showed increased hail or whatever the logarithm meant, but did not reach statistical significance for several reasons already mentioned.

If H_0 is true, the relation between x and y is random and can be replaced by other random allocations of x_i to y_j. This is systematically done by permutation of the terms in the samples x or y. Permutation changes only the covariance, the last expression in equation 2, all other terms are preserved. This condition is called "fixed marginals" for binary samples expressed in a 2 x 2 table.

There are $n!$ equally probable possibilities to rearrange the products $y_i \cdot x_j$. If all permuted R_i are sorted and plotted from both ends of smallest to larger and largest to smaller, a cdf of $min(P_i, 1 - P_i)$ is obtained. The endpoints of the cdf are the extreme correlations for both x and y sorted. The correlation between ascending x versus descending y gives the most negative or smallest R_i. Ties in x or y lead to repetitions of the same R_i and the probability increases in steps of $1/n!$. We checked numerically that the complete permutation of small binomial samples $n \leq 7$ arrives at probabilities which are *identical* to Fisher's exact solution for 2 x 2 tables.

In practice data of size $n!$ cannot be handled and the resolution 1/n! is not needed. Therefore the permutation distribution is approximated by N random samples. This is called resampling, rerandomization or Monte Carlo method. Such a plot starts and ends at $P = 1/N$. The blue curve in Figure 3 shows the approximation by N=100'000 points. Each permutation is represented by a point. The points are connected to a line zigzagging from 0.001% to 0.002%, 0.003% and so forth. From 0.1% onward the curve becomes stable as may be seen.

The precision in terms of the std of P is given by

$$\sigma_P = ((P - P^2)/N)^{0.5} \tag{7}$$

This is also found in [12, p. 97]. The resampling is done with replacement. The consequences of replacement are negligible in the context of permutations. It just means that the complete permutation distribution is never met exactly, even when N is equal or larger than $n!$, but the error is known.

Another combinative method to calculate probabilities is bootstrapping, mostly used for confidence intervals [6]. The bootstrap creates new samples by selecting n times from y, from x or from both, with replacement. In this way an association between y and x is also broken. A bootstrap without replacement is like a permutation. Bootstrapping with replacement creates new samples with different mean and std. We bootstrap y, the most critical distribution. This simulates new thunderstorms as if the question was what would happen if the experiment was repeated. The red curve in Figure 3 shows the result for applying bootstrap to E_{GR} of the 83 hail days 100'000 times. The coincidence of the red curve with the blue curve from permutation is most remarkable. The probabilities are 0.38% for both. Doing the same for the 253 cells shows also good coincidence (0.38% for permutation and 0.34% for bootstrap). The difference between permutation and bootstrap in Figure 3 is negligible. In certain conditions the differences could be considerable as explained in the appendix A. But in the case of the hail data, distributions and correlations are not sensitive to the model of calculation applied. Otherwise detailed knowledge of the experimental circumstances may have been necessary to chose the adequate model, if possible.

Calculations based on R and permutation have a great advantage insofar as the probabilities $P(R|H_0)$, $P(dif|H_0)$ and $P(rr|H_0)$ are identical because sorting the permutations of these parameters form the same succession. Not so bootstrap where the bootstrapped variables obviously change. Even the seemingly simpler models using averages *avs* and *avn* are more complicated because permutations

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 July 2021 doi:10.20944/preprints202107.0666.v1

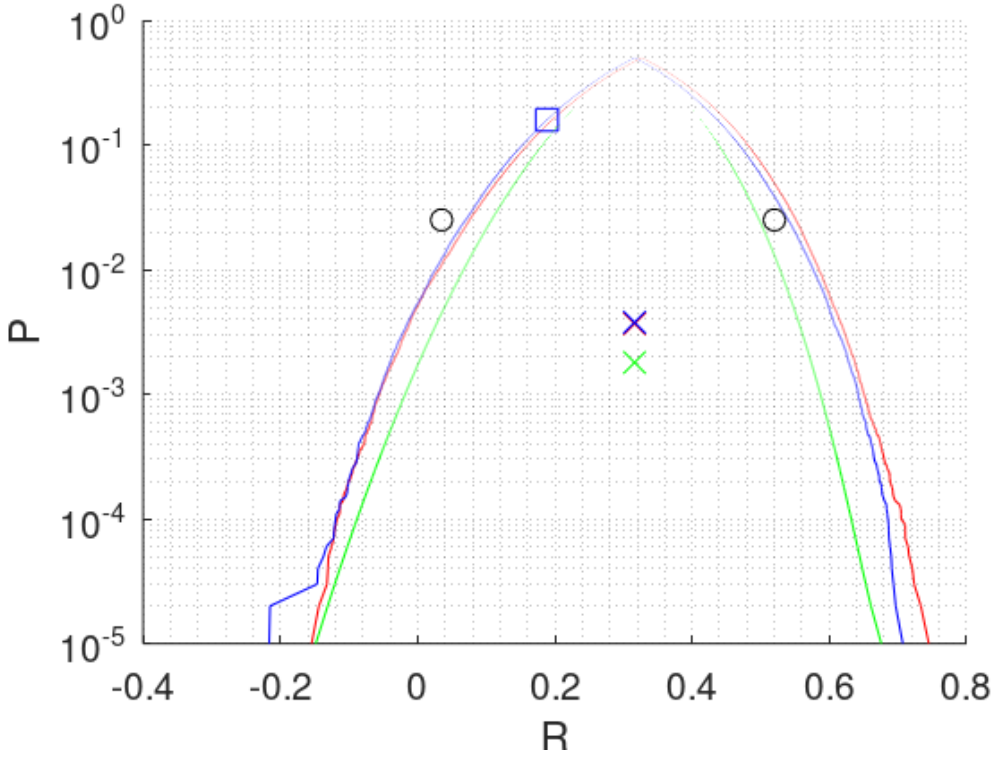

Figure 4. Three curves cdf of $min(P, 1-P)$ to read off CI for the methods based on Fisher's z (green), permutation CIP (blue) and bootstrap (red). The two black circles indicate the CI obtained by BCa bootstrapping. The crosses remind $P(H_0)$ of 3. The blue square indicates R at a probability of 15.9%, see text. The underlying y are the hail energies of the 83 days, x the seeding coverage.

change several components, namely $sum(y. * x)/sum(x)$ and $sum(y(x = 0))/length(y(x = 0))$. All these latter models calculate different probabilities for R, dif and rr as seen in Table 1.

In the further course of this work the regression model is pursued. The other data in Table 1 are less compact, but all in a range of probabilities far below 2.5%. This is good evidence for a statistically significant correlation between E_{GR} and sc in the sense that the hail energy is increased when seeding.

3.3. Confidence intervals and standard error

The next question is about the accuracy of R and the derived dif and rr. Confidence intervals (CI) are the means to treat these issues. Resampled distributions with R_i are needed assuming the alternative hypothesis H_1 that R found in the experiment is true and should correspond to the median of the resampled R_i. An old solution for normally distributed y and x is again Fisher's z for CI.

$$z_i = 0.5 \cdot (n-3)^{-0.5} \cdot \ln\left((1+R_i) \cdot (1-R)/((1-R_i) \cdot (1+R))\right) \tag{8}$$

As above, a standard normal distribution with z_i is expected when y and x are Gaussian. The green curve in Figure 4 is again shown for comparison with the solutions by combinative methods.

Efron [6] proposed "bivariate" bootstrapping to calculate CI by resampling the originally associated x_i and y_i pairwise with replacement. In this way the correlation of the sample is preserved in the average of all bootstraps producing R_i, although median$(R_i) = R$ is not guaranteed. Performing this bootstrap leads to the red curve in Figure 4.

Instead, permutation keeps all terms of y and x but varies the associations between, which destroys any correlation. In the course of this work a simple and transparent way was found to impose the observed (or any other possible) R as the median of all permutations. After permutation

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 July 2021 doi:10.20944/preprints202107.0666.v1

a random sequence of length m_1, is sorted to produce the maximum positive or, when R is negative, the maximum negative correlation. This m_1 is used to compensate, by construction, for the loss of correlation in the permuted terms. The task is to find the correct m_1 which guarantees median$(R_i) = R$. An adequate m_1 to start with is:

$$m_1 = R \cdot (n - m_0)/R_{max} \tag{9}$$

The term m_0 is an optional number of randomly selected pairs keeping their original association (as with bootstrapping). The portion of the sample subjected to permutation is $n - m_0 - m_1$. This procedure to resample R_i by permuting and maximising the association of m_1 random terms may be named "correlation imposed permutation" (CIP). CIP keeps all terms of y and x and plays with the associations between y and x to form a permutation distribution for CI. There are $n!/(m_1! \cdot m_0!)$ permutations, approximated by N terms R_i as explained in 3.2.

Equation 9 situates the median of the permutations already in the vicinity of R. The correction to establish a better m_1 for the next approximation is (R-median$(R_i)) \cdot (n - m_0)/Rmax$. By two or three further runs median$(R_i) = R$ is reached with adequate precision.

Figure 4 shows the blue curve for CIP, n = 83 days, $m_0 = 0$, $m_1 = 32.7$. A non integer m_1 is needed for the accuracy of the condition median$(R_i) = R$. It is realized by alternating in the present case between 7 times $m_1 = 33$ and 3 times $m_1 = 32$. The blue curve in Figure 4 is close to the red curve as already found in Figure 3. Again, the hail data are indifferent with respect to the two models applied for calculation.

Concerning m_0 there is an interesting suggestion based on equation 8: $P(R|H_0)$, indicated by a green cross in the middle of Figure 4, is identical to $P(R_i = 0|H_1)$. As an option, this condition could be applied to determine m_0 in CIP. Increasing m_0 decreases slightly CI. Introducing $m_0 = 18$ for the 83 hail days and $m_0 = 96$ for the 253 hail cells complies with this option.

In Figure 4 the confidence interval CI is the distance between the two tails of a curve at e.g. $P = 2.5\%$. At this level the CI is about four std (3.9 for normal distributions) and comprises 95% of all randomly resampled cases. Two black circles are noted outside the curves. They were calculated by the "bias corrected and accelerated" (BCa) bootstrap method going also back to Efron [13]. BCa is complicated and seems to us less convincing than the simple bootstrap or CIP. It was checked that CIP fits best Fisher's z when the samples y and x are representative for normal distributions. A survey on bootstrapping dealing also with shortcomings is found in [5]. A further critical point mentioned by Cox [14] are samples too small to be representative for a parent distribution. The appendix deals with this problem.

Instead of reading the curves for CI at 2.5% we prefer the blue square at a probability of 15.9% in Figure 4. The value of 15.9% corresponds to $R - \sigma$ in normal distributions. The standard error (σ) is an adequate measure of error. The interesting side is towards zero effect, therefore the parameter $-\sigma$ will be shown in Table 2. The other side of the 15.9% probability is asymmetric and vulnerable with respect to the parameter rr. The influence of a nearby singularity at R_{cr} may distort the cdf, see equation 5.

3.4. Re-evaluated results of Grossversuch IV

The most important results of the statistical evaluations are found in Table 1 and in Figures 3 and 4. The following Table 2 provides some further insight. It starts with the results of the regression model in rows 1 and 2, continuing with a binary x reducing sc to 0 or 1 in order to compare the present evaluations with results presented in 1986 [3].

Statistical significance is best in rows 1 and 2 because the information contained in sc is used. The bold scores show the most reliable results. Looking at cells is closer to the question asked, but the randomization was done for days. Therefore $P(H_0)$ earns more credit when calculated for days. The difference between the evaluation of $P(H_0)$ in row 1 and 2 of Table 2 is astonishingly small in view of the big difference of n. The aggregation of data from cells to days reduces stochastic variations as well as skewness and kurtosis.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 July 2021 doi:10.20944/preprints202107.0666.v1

Table 2. Results for hail data of Grossversuch IV.

Experiment		Seeded	Non-s.	$dif/cell$	$dif-\sigma$	$P(H_0)$	rr	$rr-\sigma$
E_{GR} versus sc	days	34	49	1610	987	**0.4%**	3.3	2.0
E_{GR} versus sc	cells	93	160	**1583**	966	0.4%	**3.0**	2.0
Means of two groups								
E_{GR} versus seeded, non-seeded	days	34	49	1310	710	2.0%	2.6	1.6
E_{GR} versus seeded, non-seeded	cells	93	160	1615	995	0.6%	3.1	2.0
Two groups (for comparison to [3]: cells planned for seeding but not seeded are attributed to seeded group)								
E_{GR} versus planned, non-planned	cells	113	140	1083	500	3.7%	2.2	1.4
Federer[3] Table 21, $C(\alpha)$ test	cells	113	140			1.9%	2.2	1.5
2×2 contingency table								
hail, no-hail versus seeded, non-seeded	cells	78+15	111+49	0.14	0.09	0.5%	1.2	1.1
idem, for hailpads (213 cases)	cells	45+29	64+75	0.15	0.08	2.2%	1.3	1.2

In rows 3 and 4 of Table 2 most information with respect to unsatisfactory seeding is lost. A big misinterpretation happens for row 3 because 17 non-seeded cells are taken into account as seeded in the seeded days. Only 3 cells occurring alone on 3 seeded days shift to non-seeded. Correspondingly $P(H_0)$ jumps to 2.0%. The loss in row 4 is less severe because 20 cases of planned but not performed seeding are transferred to non-seeded. Therefore the influence on $P(H_0)$ is not dramatic.

To allow a comparison with the results of [3, Table 21, last row], all 20 non-seeded cells on planned seed days were taken as perfectly seeded in rows 5 and 6. This is too much of inaccuracy leading to the loss of statistical significance in our evaluation which is more rigorous than the $C(\alpha)$-test with its assumptions.

Row 7 shows that the probability for a cell to produce hail is significantly increased by some 20% when seeding. The data from hailpads (row 8) confirm this finding, but the significance becomes marginal because of reasons discussed later. For rows 7 and 8 the results of bootstrapping were chosen because the fixed marginals anticipated by permutation are not adequate here and the differences in 2 x 2 tables notable: $P(H_0)$ = 0.7% and 2.8% would be obtained. A more impressive example is discussed in the appendix.

To sum up Table 2: Seeding increased the hail energy by a factor of 3, maybe only 2 (-1 std), the difference with respect to non-seeded was about 1600 MJ per cell and the chance to obtain this result accidentally was 0.4%, therefore statistically significant. The results hold for an average seeding of $\overline{sc} = 0.48$. An extrapolation to perfect seeding was not dared.

Not included in Table 2 are some further evaluations performed with cleaned up sets of data: either 118 cells of lifetimes less than 15 minutes for the 45 dBZ contour, or 39 cells with unsatisfactory seeding ($sc < 1/3$) could be excluded. The latter was planned in the original design of Grossversuch IV, see [3, p. 925]. The first 4 rows of Table 2 were combined with one or both of these exclusions yielding 12 further evaluations. All these evaluations show a similar picture of increase for seeding, all at a significance level below 2.5%. A trend to still lower $P(H_0)$ than in Table 2 was observed when excluding cells of short duration. All these different evaluations and models form a homogeneous picture. Even the linear regression associated with sc may be changed to a power p within $0 < p < 2$. The homogeneous picture does not change. A power p very close to zero leads to the binary simplification seeded or non-seeded. The preparation of the data and the evaluations are easily performed using the spreadsheat "DataHail-FMA.xls" available in the supplement. The evaluation of $P(R|H_0)$ in the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 July 2021 doi:10.20944/preprints202107.0666.v1

spreadsheet is based on the first four moments of the permutation distribution, a method proposed by Pitman [15]. This procedure is less robust than permutations or bootstrapping but quick and precise for the hail data. The spreadsheet contains also the calculations concerning autocorrelation, which is the next issue.

Federer [3, p. 929] observed a weak intra-day correlation, amounting to 0.33 for $\ln(E_{GR}+1)$. For non-transformed E_{GR} and non-seeded cells we found a lag-1 intra-day autocorrelation of R = 0.47 at $P(H_0)$ = 1.0%. For cells on seeded days the autocorrelation disappears: $R = -0.05$ at $P(H_0)$ = 56%. As E_{GR} changes under the influence of a variable *sc*, the autocorrelation is destroyed. R, *dif* or *rr* are not affected by the autocorrelation, only the calculation of $P(H_0)$ may be too optimistic when the independence of the units is not perfect. The following experiment localizes the effect of autocorrelation on $P(H_0)$.

The distribution of E_{GR} with cells is varied in the set of non-seeded data, while the daily total of E_{GR} remains unchanged. The two most extreme cases are:

1. Each cell contributes the same amount to the daily E_{GR} of non-seeded cells, corresponding to total intraday autocorrelation. The result of the permutation test for the 253 cells is *rr* = 3.0, $P(H_0)$ = 0.27%.
2. The daily total comes from only one cell, the other cells of the same day are without hail. In this case *rr* = 3.0, $P(H_0)$ = 0.74% is obtained.

In this bandwidth from 0.27% to 0.74% the observed result is found: *rr* = 3.0, $P(H_0)$ = 0.38%, equal to the result for days (*rr* = 3.3, $P(H_0)$ = 0.38%). It seems that the intraday autocorrelation of non-seeded cells is not really disturbing. Also Federer [3, Table 22] based 16 of their 21 tests on cells. Autocorrelation could have been a real problem if the several severe hailstorms would have been aggregated on a few days. But there is only one day, 18 July 1978, non-seeded, with two very large cells, causing the daily maximum of $E_{GR} = 43000MJ$. A plausible explanation of the autocorrelation is the aggregation of cases with zero or little hail on days with meteorological conditions not suitable to produce severe storms.

Autocorrelation is not observed in the data from hailpads. This has to do with an interesting question: how do stochastic uncertainties in the measurements influence the results? From [7] and [8] we estimate the uncertainty of the radar based E_{GR} at 20 to 25%. In case of a systematic multiplicative error in the radar calibration and thus in E_{GR} the significance level $P(H_0)$ remains unchanged. The reason is that linear transformations do not change R. If the error in E_{GR} is stochastic, it has an impact on $P(H_0)$, as a simple numerical test can show. We added a random error of 20% to E_{GR} of unit days, first row in Table 2, repeating the experiment 100 times. The significance level diminishes as $P(H_0)$ increases from 0.33% to an average of 0.55%. When increasing the error to 40%, there is a further impairment of $P(H_0)$ to 1.1%. In both cases *rr* remains practically unchanged and $P(H_0)$ remains below 2.5%.

We learn from this that data suffering from too much inaccuracy loose power. Unfortunately this seems to be the case for the data obtained from hailpads. Also the hailpad data show an increase of hail energy for seeded data, but statistical significance is not reached, e. g. for cells *rr* = 1.58, $P(H_0)$ = 16%. Federer's Table 21 reports *rr* = 1.58, $P(H_0)$ = 24% for the $C(\alpha)$ test. The data from hailpads lack 40 cells mainly from the year 1982. But this can not be the decisive point, as the radar data reach for the same 213 cells still *rr* = 2.79, $P(H_0]$ = 0.7%. We suspect that the sampling by hailpads introduces intolerable stochastic variations and we tested this hypothesis by looking at the intraday autocorrelation of the hail energies from hailpads for unseeded days. Comparing the radar data for the same 79 cells to the hailpad data reveals R = 0.47, $P(H_0)$ = 1.2% for the radar, degrading to R = 0.09, $P(H_0)$ = 16% for the hailpads. This is a strong hint that the accuracy of the hailpad measurements is not adequate to show the intraday autocorrelation. Furthermore, the total hail energy is 0.41 times that of the corresponding E_{GR}. The conjecture is that the hailpad network not only introduces large stochastic errors, but looses information which is important for the evaluation of hail energies. Less demanding is the question

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 July 2021 doi:10.20944/preprints202107.0666.v1

whether at least one hailpad was hit, indicating hail or no hail. Again, the hailpads identified less hail (51%) than the radar (75%) for both seeded and non-seeded experimental cells.

4. Discussion

4.1. What transformations do

We criticized the 1986 evaluation of Grossversuch IV in [3] on behalf of the logarithmic transformation of the hail energy E_{GR}. But something positive about it was already mentioned: As the influence of a few dominating cases is transformed away by the logarithm, the chance is increased to find a significant correlation between $\ln(1 + E_{GR})$ and *sc* or its binary transformation. Indeed this correlation is positive at a level $P(H_0)$ = 0.008% for *sc* and 0.006% for the binary $x = 0$ or 1. This may help to rule out H_0 and confirm the result that seeding increased hail.

It could have been quite different. We give an example which leads to the situation of a positive correlation for E_{GR} and a negative correlation for $\ln(1 + E_{GR})$, both at very low levels of $P(H_0)$. A simple synthetic example works with three different values of E_{GR} and $\ln(1 + E_{GR})$, respectively:

- 100 seeded and 20 non-seeded cells with both E_{GR} and $\ln(1 + E_{GR}) = 0$
- 100 non-seeded cells with E_{GR} =150, $\ln(1 + E_{GR}) = 5$
- 20 seeded cells with E_{GR} = 3000, $\ln(1 + E_{GR}) = 8$

For E_{GR} the correlation is positive, $P(H_0)$ = 0.01%, *rr* = 4.0. For $\ln(1 + E_{GR})$ the correlation is negative, $P(H_0) \ll 0.001\%$, *rr* = 0.32. A Wilcoxon-test where the data are filled into three ranks 0, 1 and 2 yields also a negative correlation, $P(H_0) \ll 0.001\%$, *rr* = 0.40. These two transformations invert a statistically significant result to the opposite with even stronger significance. This example is a simplified version of what happens if seeding would prevent hail in small storms and enhance hail growth in a few very large storms. This scenario is not unrealistic and, if it should happen, hard to prove. In fact, the data of Grossversuch IV are close to this pattern, but only the increase of E_{GR} for seeded storms reach statistical significance.

It is interesting to see how the results of the 1986 study react on transformations to the logarithm or to ranks (WMW-test, after Wilcoxon, Mann and Whitney). Without transformation we find in [3] Table 22 always $rr > 1$, namely $1.26 < rr < 4.38$. The transformations produce both $rr > 1$ and more often $rr < 1$. The latter reminds our artificial example. Our study does not speak against hail suppression in short lived cells (life time less than 15 minutes of radar reflectivity >45 dBZ). We find $rr < 1$, but this could be real as well as accidental.

The point is that the parameters *R*, *dif* or *rr* should be calculated using the variables *y* and *x* describing the issue, see [16] or [17, p. 246 ff.]. Nonlinear transformations of *x* or *y* lead away from what was asked. This criticism holds also for the transformation to ranks. Something else is transforming *R* to students *t* or Fisher's *Z*. This is just changing the functions to calculate the probability (which for non-normal data is no more a *t* or a normal distribution as permutations show). The condition is, that sorting R_i or sorting a transformation of R_i keep the same sequence when using permutations.

4.2. Multiplicity effects

The authors of the 1986 paper state about the unfavorable seeding effect shown in their Table 21, that statistical significance "may easily be attributed to the multiplicity effect (which means that some out of a number of tests turn out significant by pure chance), but seeding influences are also a possible explanation" [3, p. 949].

The question of the study was formulated as follows [3, p. 923]: "Do the experimental cells on seed days and no-seed days differ in the response variable in a statistically significant way?" As this question allows for either increasing or decreasing effects of seeding on hail formation, the significance level of the usual 5% is split into 2.5% for positive and 2.5% for negative influence. The planned evaluation of $\ln(E_{GR} + 1) - f$ in [3] missed the goal of finding out, whether hail suppression worked in the

 doi:10.20944/preprints202107.0666.v1

important severe storms, and the logarithmic a posteriori predictor f added a disturbing complexity. The present analysis remains as close as possible to the original question of the study and keeps the hail energy E_{GR} as response variable. Furthermore it makes use of the information on rudimentary or non-seeding when seeding was planned. The permutation test of the correlation between E_{GR} and sc in agreement with the bootstrap is the most adequate mathematical treatment for the question asked. Therefore it earns due credit and does not fall into the category where "multiplicity effect" [3, p. 949] could happen.

4.3. Possible mechanisms

Theories or modelling of the cloud physical processes increasing the hail energy of seeded storms are outside the goal of this paper. But some ideas are put forward to show that the observed result is plausible. The formation of hail in a thunderstorm is a fortuitous matter. Ice and super-cooled water are necessary as well as updrafts matching the fall-speed of hailstones, keeping them half an hour in zones of super-cooled water and enabling several up and downs. Complicated non-linear processes are involved which implies high sensitivity to small differences in the initial state of the atmosphere. No wonder that the prediction of hail energy is difficult. Seeding with silver iodide triggers freezing of supercooled water at temperatures up to -5 degrees Celsius. This may reduce hail due to competition, but, it may as well enable new scenarios for the formation of hail that would not have been possible without seeding.

Hail forms when super-cooled liquid water is captured by ice particles and then freezes on the surface. The processes and variables involved are complex, see for instance [18,19] and [20]. The primary material for growth is super-cooled water present in the form of droplets. We can distinguish between two scenarios depending on the balance between the amount of super-cooled water and the number of ice particles including hail embryos and hailstones.

In the first scenario super-cooled water is abundant, even after seeding, and there are not enough ice particles to deplete the supercooled water substantially. Seeding will create ice crystals and increase their number in zones up to -5 C, leading to more hail embryos also in places where otherwise no ice would have been generated by natural ice forming nuclei. This gives way to more intensive as well as new scenarios of hail growth starting in relatively warm zones. Given abundant supercooled water, this forms the basis for the growth of additional and larger hailstones. Without sufficient competition, the opposite happens of what seeding is supposed to do.

In the second scenario there is a shortage of super-cooled water in relation to the number of ice particles. By increasing the number of freezing nuclei seeding will enhance competition among embryos. This may inhibit the growth of large hailstones, which is the underlying assumption of hail suppression by seeding.

The results presented in this study, however, suggest that the first scenario is dominating, at least for the severe storms which add up to a large part of the total hail energy. This is supported by recent findings that dry growth is unimportant for large hail [19]. Examples for large hailstones indicating wet growth are given in [21]. Another example is the 766 g hailstone of Coffeyville (NCAR Fact Sheet, October 1970) with typical protrusions indicating wet growth. These examples stand for severe storms with plenty of super-cooled water. It is doubtful whether seeding can reduce the amount of super-cooled water adequately. In any case, some of the extra ice particles produced by seeding may stick to the wet surface of growing hailstones, which would enhance growth and counteract the competition theory of seeding. These are plausible explanations of how seeding could enhance hail.

Last but not least an increase of the number of hail-cells was found when seeding: rr = 1.2, $P(H_0)$ = 0.5% and 0.7% for bootstrap and permutation, respectively. When looking at the days as unit no such increase is observed. The interpretation is that some experimental days offered just unsuitable conditions for hail, whether seeded or not. The observed intra-day autocorrelation supports this suggestion. On the other hand, the triggering of supplementary hail cells by seeding on days that have already produced hail can not be detected when analyzing days.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 July 2021 doi:10.20944/preprints202107.0666.v1

A last question concerns the factor of 3 found for the increase of hail energy when seeding. It seems large. Is it due to a larger number or size of hailstones or to a larger area? A relatively modest factor of 1.2 can be attributed to an increased probability that seeded cells produce hail, as documented in Table 2, row 7 and 8. More important is the factor of 1.8 found in Table 13 of [3] for an increase of the area touched by hail when seeding was planned (two-sided $P(H_0) = 2.9\%$, $C(\alpha)$-test). Most probably the factor 1.8 underestimates the reality because *sc* was replaced by what was planned. We expect that an analysis using *sc* would show an increase somewhere between 2 and 3, as well as better statistical significance, similar to the differences found for E_{GR} in Table 2, comparing row 2 with row 5. Unfortunately the data of the hailed area for the individual cells are not contained in the data given by [3]. Anyhow, the statistical treatment of the question concerning the area is the least demanding because it boils down to the number of hailpads touched by hail. In this respect the density of the network may yield sufficient resolution.

The considerably increased area of hail as well as the increased probability for hail sustains the idea that seeding enables additional hail scenarios. May be that seeding created also some situations favourable to grow larger hailstones. As the energy E_{GR} is proportional to D^4, only a small difference of hailstone diameter has a large impact on hail energy. On the other hand, the size of the largest hailstones depends on the updraft velocity which is governed by the dynamics of the storms.

5. Conclusions

The conclusion of the present re-evaluation is, that the seeding in Grossversuch IV *increased* the hail energy by $dif = 1600$ MJ/cell, which is a factor of $rr = 3$. This pertains to an average seeding of $\overline{sc} = 0.48$. The precision is rather marginal, $rr = 2$ is within one std. But the statistical significance is almost sure, as all evaluations yield $P(H_0)$ below 2.5% and these using the available information about the quality of seeding are nearly an order of magnitude below 2.5%, see Table 2, row 1 and 2.

From a physical point of view the result is not unrealistic, although a model proving that it must be so can not be given at this time. Most likely seeding enhances further scenarios of hail production starting at warmer temperatures, increasing the area of hailfall.

Stochastic variations reduce statistical significance. The hailpad network, although it was one of the most dense and expensive we know of, was not good enough to measure reliably the total hail energy. At least it revealed that the area touched by hail when seeding was enlarged by a factor of 1.8 according to [3] and even more when some inaccuracy concerning the seeding is removed.

The statistical evaluation required much space because the 1986 study [3] was not satisfactory in this respect and some problems associated with asymmetric or heavy tailed distributions as well as non representative samples are still a challenge. To go round these problems by applying non-linear transformations to the raw data such as a logarithmic transformation or a conversion to ranks inserts a distortion between the question and the answer. This was disturbing in the original evaluation of Grossversuch IV.

Models are needed to calculate probabilities. Difficulties may arise from more or less clear assumptions underlying a model. Permutation and bootstrap used here are quite transparent. But sometimes the data alone are not sufficient to find the correct probabilities as in the contingency table $\left|\begin{smallmatrix} 6 & 0 \\ 0 & 2 \end{smallmatrix}\right|$ when the condition of fixed marginals is not correct, see appendix.

The sample size n divided by the kurtosis of the sample is an indicator of the effective sample size. If not much larger than 1 it is indicative of a outlier problem leading to differences between permutation and bootstrap as explained in the appendix. It is recommended to calculate $P(H_0)$ by permutation as well as by bootstrap. If there is agreement, the sample is indifferent with respect to these models. The continuation may be permutation and regression which offers a compact solution for the parameters R, dif and rr. The present work opened up a way to evaluate also CI by permutation CIP. This deserves further attention.

Funding: This research received no external funding

Acknowledgments: Matthias Auf der Maur helped to introduce Octave. William Duddlestone is acknowledged for hints and linguistic improvements.

Author Contributions: The statistical calculations and CIP have been developed by the first author. With reference to the age of Armin Auf der Maur, the young colleague Urs Germann is appointed as communicating author. He contributed to the conception of the paper and he was a critical reviewer of the statistical part. He contributed to the parts about Grossversuch IV, the measurements by radar and some possible mechanisms for the observed effect of seeding.

Conflicts of Interest: The authors declare no conflict of interest.

Supplementary Materials: Programs in Octave, compatible with Matlab, to calculate the cdf's of $P(H_0)$ and $P(H_1)$, together with the data files, are found in the supplement http://www.mdpi.com//xx/1/5/1.

Appendix A. Modeling an experiment by permutation or bootstrap

A notable passage is found in DiCiccio and Efron [13, p. 191]: "In most problems and for most parameters there will not exist exact confidence intervals." The problem is that the exact model to calculate the probabilities is rarely available. Even for the simpler calculation of H_0 an example for possible difficulties will be given later.

The differences between the two models permutation and bootstrap can be demonstrated best by using an example with an extreme outlier in y. Permutation creates R_i containing the outlier just once. Bootstrap, instead, varies its appearance between none (in about 37% of the draws), once (37%), twice (18%) and more times (8%). This must lead to a difference between the permutation and the bootstrap distribution. An additional difference appears in the calculation of CI as permutation associates an outlier in y with all terms of x whereas the bivariate bootstrap keeps the outlier always together with its originally accompanying term.

The most extreme outlier appears in a sample of $n-1$ equal and one divergent value. Such a sample has the largest possible standardized moments of order $k \geq 3$: $\beta_k = m_k \cdot m_2^{-k/2}$ (m_k is the central moment of order k). The proof is simple as any change to the extreme sample leads to less extreme moments β_k. It is readily calculated for a sample [1, 0, 0, ..., 0]:

$$\beta_4 = n - 2 + 1/(n-1) \tag{A1}$$

We use here the kurtosis β_4 rather than the skewness β_3 because it indicates symmetric as well as asymmetric heavy tailed samples. Furthermore $\beta_4 \geq (\beta_3)^2 + 1$ holds, see e.g. [5].

As β_4 reaches quasi n for an extreme outlier it suggests itself to use n/β_4 as an indicator for the number of effective terms. The less important terms are those in the bulk of the distribution. The smallest possible $\beta_4 = 1$ is realized by a symmetrical binary sample.

If n/β_4 is about 1 or 2, the sample is characterized by just one or two prominent outliers. Such a sample is not representative because it can be only loosely associated with a parent distribution. This issue was mentioned by Cox [14]. The hail data E_{GR} for days as well as for cells range close to $n/\beta_4 = 8$. Eight seems sufficient not to prevent agreement between permutation and bootstrap as Figures 3 and 4 suggest. The issue n/β_4 deserves further attention.

A small n/β_4 is not the only problem. Discontinuities due to ties and fixed or not fixed marginals can provoke difficulties. An impressive example is the 2 x 2 table $\left|\begin{smallmatrix} 4 & 0 \\ 0 & 4 \end{smallmatrix}\right|$. It has the smallest possible $\beta_4 = 1$ for both y and x. It stands for Fisher's famous experiment with a lady who successfully detects the four cups where the milk was added after the tea and the other four cups where the milk was poured in first, see [17, p. 59]. The blue staircase in Figure A1 obtained by permutation or by Fisher's exact solution models exactly the case when the lady is informed that there are four cups of each kind. This leads to fixed marginals, restricting the possibilities for hits and faults to 0, 2, 4, 6 or 8 in 70 equally probable arrangements. Permutation keeps to this scheme and yields the correct result $P(H_0) = 1.4\%$. Bootstrap, on the other hand, comes up with the red points in Figure A1. It describes a more sophisticated experiment: The partition of the eight cups is no longer fixed and not known to the lady. Fixed marginals are abolished and there are now 9 possibilities for hits and faults in 254 different

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 July 2021 doi:10.20944/preprints202107.0666.v1

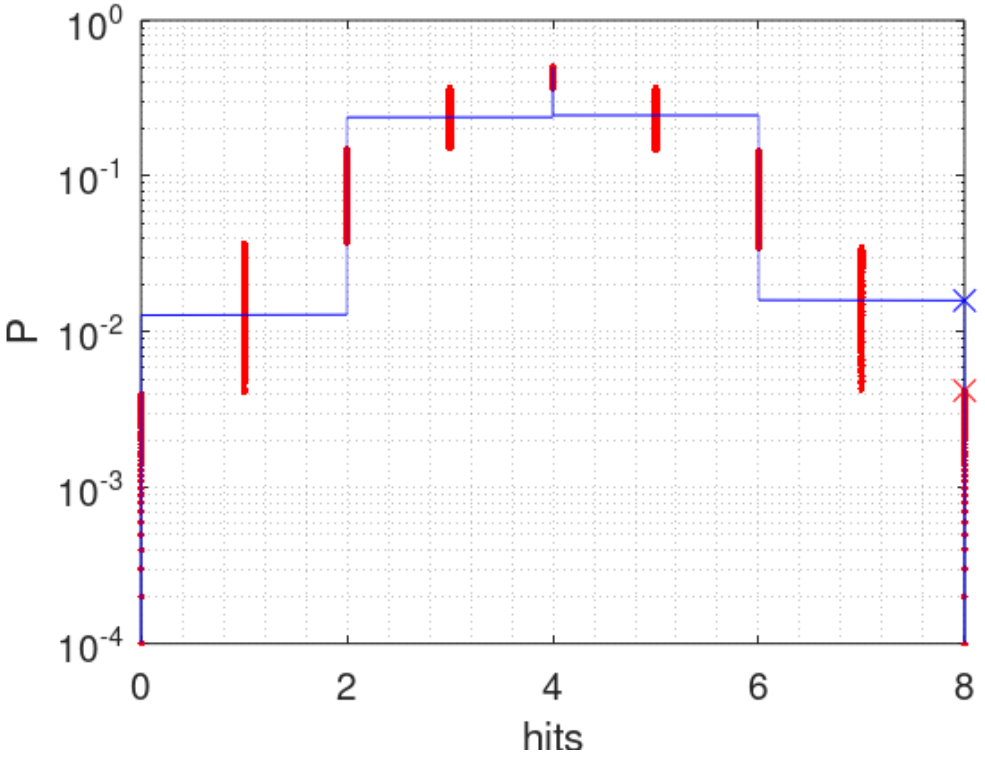

Figure A1. Eight cups of tea tasted by the lady in Fisher's experiment (see [17], p. 59). The probability for accidental hits assuming H_0 true is calculated either by permutation (blue) or by bootstrap (red). The blue cross indicates the statistical significance when the partition is known, the red cross when not known and everything is possible.

arrangements. The statistical significance for the correct answer of the lady is therefore $P(H_0) = 0.4\%$. Assume now that the experimenter or tossing a coin decided for two cups with milk added to the tea (= 1) and six cups with milk poured in first (= 0). The contingency table of the correct guess is $\left|\begin{smallmatrix} 2 & 0 \\ 0 & 6 \end{smallmatrix}\right|$. However, bootstrapping these data would yield $P(H_0) = 1.1\%$, permutation $P(H_0) = 3.6\%$, whereas $P(H_0) = 0.4\%$ is correct. A sample with four 0 and four 1 must be bootstrapped to obtain the red points in Figure A1. Tossing a coin delivers this favourable condition in only 28% of the trials. This example illustrates certain limitations when the observed samples are the only source of information. Furthermore the table $\left|\begin{smallmatrix} 2 & 0 \\ 0 & 6 \end{smallmatrix}\right|$ is close to a sample with two outliers, which is a warning.

Contrary to the pitfalls described above, samples rely often on many similarly important values, leading to a large $n/\beta_4 > 10$. As a consequence the differences between permutation and bootstrap are expected to vanish, at least in the region of the interesting P values, maybe not near the end of the tails where $P = 1/N$. This is found for normal distributions ($n/\beta_4 \approx n/3$), but also for the hail data as Figures 3 and 4 show. For both 83 days and 253 cells $n/\beta_4 \approx 8$. When permutation and bootstrap yield compatible results, they earn confidence. Ultimate precision is seldom possible and not required.

Programming the presented methods in Octave, Matlab, R, Python or any other similar language one is familiar with is not difficult. To preserve the association between y and x in bivariate bootstrapping or permutations with $m_0 \geq 0$, y and x are packed into a complex vector. In a for- or do-loop the permutations or bootstraps are executed N times by the Octave command "y(randperm(n))" or "experimental_rnd(y)", respectively. N = 10'000 is quick, N = 100'000 provides the intended precision, needing on a modern laptop with intel CORE i7 about one minute. Data and codes for Octave are found in the supplement. The calculation of BCa follows a blog by methodsconsultants.com/posts/understanding-bootstrap ...r-boot-package by J. Albright, 2019.

References

1. Wieringa, J.; Holleman, I. If cannons cannot fight hail, what else? *Meteor. Z.* **2006**, *15*, 659–669.
2. Rivera, J.A.; Otero, F.; Naranjo Tamayo, E.; Silva, M. Sixty Years of Hail Suppression Activities in Mendoza, Argentina: Uncertainties, Gaps in Knowledge and Future Perspectives. *Frontiers in Environmental Science* **2020**, *8*, 45. doi:10.3389/fenvs.2020.00045.
3. Federer, B.; Waldvogel, A.; Schmid, W.; Schiesser, H.H.; Hampel, F.; Schweingruber, M.; Stahel, W.; Bader, J.; Mezeix, J.F.; Doras, N.; DAubigny, G.; DerMegreditchian, G.; Vento, D. Main results of Grossversuch IV. *J. Appl. Meteor. Climatol.* **1986**, *25*, 917–957.
4. Sulakvelidze, G.K.; Kiziriya, B.I.; Tsykunov, V.V. Progress of Hail Suppression Work in the USSR. In *Weather and Climate Modification*; Hess, W.N., Ed.; Wiley, 1974; pp. 410–431.
5. Bishara, A.J.; Hittner, J.B. Confidence intervals for correlations when data are not normal. *Behaviour Res. Meth.* **2017**, *49*, 294–309.
6. Efron, B. Bootstrap methods: Another look at the jackknive. *Annals of Statistics* **1979**, *7*, 1–26.
7. Waldvogel, A.; Schmid, W.; Federer, B. The Kinetic Energy of Hailfalls. Part I: Hailstone spectra. *JAM* **1978**, *17*, 515–520.
8. Waldvogel, A.; Federer, B.; Schmid, W.; Mezeix, J.F. The Kinetic Energy of Hailfalls. Part II: Radar and Hailpads. *JAM* **1978**, *17*, 1680–1693.
9. Waldvogel, A.; Schmid, W. The Kinetic Energy of Hailfalls. Part III: Sampling Errors Inferred from Radar Data. *JAM* **1982**, *21*, 1228–1238.
10. Schmid, W.; Schiesser, H.H.; Waldvogel, A. The Kinetic Energy of Hailfalls. Part IV: Patterns of Hailpad and Radar Data. *JAM* **1992**, *31*, 1165–1178.
11. Berry, K.J.; Mielke, Jr., P.W.; Mielke, H.W. The Fisher-Pitman permutation test: An attractive alternative to the F test. *Psychol. Rep.* **2002**, *90*, 495–502.
12. Lee, W.C.; Rodgers, J.L. Bootstrapping correlation coefficients using univariate and bivariate sampling. *Psychological Methods* **1998**, *3*, 91–103.
13. DiCiccio, T.J.; Efron, B. Bootstrap confidence intervals. *Statistical Science* **1996**, *3*, 189–228.
14. Cox, N.J. Speaking Stata: The limits of sample skewness and kurtosis. *Stata Journal* **2010**, *10*, 482–495.
15. Pitman, E.J.G. Significance tests which may be applied to samples from any populations. II. The correlation coefficient test. *J. Roy. Stat. Soc. Suppl.* **1937**, *4*, 225–232.
16. Feinstein, A.R. Clinical Biostatistics XXIII: The role of randomization in sampling, testing, allocation and credulous idolatry (Part 2). *Clin. Pharmacol. Ther.* **1973**, *14*, 898–915.
17. Berry, K.J.; Johnston, J.E.; Mielke, Jr., P.W. *A Chronicle of Permutation Statistical Methods*, 1 ed.; Springer International Publishing, 2014; p. 517.
18. List, R. New Hailstone Physics. Part I: Heat and Mass Transfer (HMT) and Growth. *JAS* **2014**, *71*, 1508–1520.
19. List, R. New Hailstone Physics. Part II: Interaction of the Variables. *JAS* **2014**, *71*, 2114–2129.
20. Aufdermaur, A.; Joss, J. A wind tunnel investigation on the local heat transfer from a sphere, including the influence of turbulence and roughness. *Zeitschrift für Angewandte Mathematik und Physik* **1967**, *18*, 852–866.
21. Levi, L.; Achaval, E.; Aufdermaur, A.N. Crystal Orientation in a Wet Growth Hailstone. *JAS* **1970**, *23*, 512–513.