The giant impact hypothesis: past, present (and future?)

William K. Hartmann
Planetary Science Institute, 1700 East Fort Lowell Road, Suite 106, Tucson, AZ 85719-2395, USA

At the request of editors, this paper offers a historical review of early work on the giant impact hypothesis, as well as comments on new data. The author hereby claims (whether believable or not) that his interest is to move towards a correct model of lunar origin, not to defend a possibly incorrect idea, just because of being a co-author of a relevant early paper. Nonetheless, the 1974 giant impact hypothesis appears still to be viable.

1. Early history

(a) Presentation of the modern giant impact model

At the request of the organizers of the Royal Society Discussion Meeting on the ‘Origin of the Moon’, I begin with a historical perspective on the current issues. Dynamicist Donald R. Davis and I, as fledgling PhD’s, gave the first formal presentation of the modern giant impact hypothesis at an International Astronomical Union Symposium on planetary satellites, held at Cornell University, Ithaca, NY, USA, 18–21 August 1974 [1]. Davis, who had recently helped bring Apollo 13 back from the Moon, had joined the Planetary Science Institute and was pioneering models of planet growth. With an eye to lunar origin, we estimated sizes to which the second largest, tertiary, etc. planetesimals in a planet’s orbital vicinity might have grown, and hence the size of the largest impact that might be experienced by a given planet. Stochastic vagaries of the second largest body’s survival in the face of various fates (such as catastrophic disruption by a smaller planetesimal or gravitational scattering out of the Solar System altogether) implied a variety of possible outcomes. We found that Moon-to-Mars-sized bodies accreted as Earth approached its present size, and might plausibly have hit proto-Earth. While presenting our work at Cornell, I emphasized that if the largest body to hit Earth were large enough, it might blow mantle material out of
Earth, which could form a ring of debris and lead to formation of a moon with few volatiles and without a large iron core, assuming Earth’s iron had already been sequestered in Earth’s iron core. Such a model seemed to explain the Moon’s lack of iron, which bedevilled lunar origin theories of that era (although, at that time, it was not generally conceded that Earth’s core could have already formed by the time of such an impact).

At the end of that presentation at the Cornell meeting, a hand was raised by A. G. W. Cameron, regarded as a prime authority on planetary growth. Cameron remarked that he and his post-doctoral scholar, Bill Ward, had been working on a similar idea of lunar origin, but from the point of view of satisfying angular momentum constraints. From that point of view, they estimated that the putative impactor might have been as large as Mars. Hartmann and Davis published an abstract in the Cornell programme book [1] and published a full paper in *Icarus* the next year [2], and Cameron and Ward published their work in a Lunar and Planetary Science Conference (LPSC) abstract the following year [3]. The giant impact hypothesis was off and running.

Based on Cameron’s comments on our talk at Cornell, Hartmann, Davis, Cameron and Ward, and the lunar community at large, have generally agreed that this ‘modern’ version of the giant impact idea was conceived independently and more or less simultaneously by both groups, so we regard issues of priority as moot. After the Royal Society Discussion Meeting in 2013, however, *Science* published a review of the meeting [4], which attributed the giant impact idea only to Hartmann & Davis [2] and, in the interest of historical accuracy, Melosh *et al.* [5] responded with a correction. Their wording, however, may have added new confusion to the historical record. They stated, ‘In actuality, two groups developed this idea contemporaneously and both discussed essentially the same idea at a Cornell conference in 1974 at which all four researchers were present’ ([5], p. 1445). Melosh *et al.* added that Hartmann and Davis ‘directly acknowledged this in a footnote in [the Hartmann/Davis 1975] paper’ and that ‘Cameron and Ward’s work appeared in early [1976]’. This wording may imply to many readers that each group presented a lunar origin paper at the 1974 conference, and the two groups then published comparable papers. As indicated in the programme book for the Cornell meeting, however, Cameron’s only formal talk (earlier in the meeting) was on outer planets, and he presented no abstract on that talk. His lunar origin remarks were only comments from the floor after my presentation. Thus, neither he nor Ward had published abstracts or papers on the topic by the time our 1975 paper was submitted. We wanted to alert the community that Cameron and Ward had reached a similar conclusion to our own about a big impact, however, and since their 1976 LPSC abstract [3] was not yet published, we included a footnote in our paper to mention their work.

Ward & Cameron [6] and Thompson & Stevenson [7] began analysis of the evolution of the putative debris disc during the next decade, but otherwise the giant impact hypothesis attracted little support in its first decade. I should probably have presented those ideas at several more major meetings but, instead, I moved on to participation in the Mariner 9 Mars mission. The hypothesis ‘blossomed’ only after the Kona conference in 1984 (see §1d).

(b) An earlier suggestion of giant impact: the ‘lost theory’ of Daly [8]

Some years after the giant impact idea became popular at the 1984 Kona conference, Don Davis and I began to receive enquiries about an earlier version of the giant impact idea—a 1946 paper we had never heard of, by a geologist named Reginald A. Daly. It was eventually discussed by Baldwin and Wilhelms in 1992 [9] and was described (without formal bibliographic reference) in a popular 2003 book about the giant impact hypothesis, as the ‘debut of . . . what has become known . . . as the “giant impact hypothesis”’ [10].

Daly (1871–1957) was a Canadian–American geologist who taught primarily at Harvard University, Cambridge, MA, USA, and was known particularly for his 1903 ideas about igneous rocks and why magma rises from Earth’s interior to its surface. He retired in 1942, and in the next few years became interested in the Moon. Daly’s lunar origin paper was a 15-page article published in the *Proceedings of the American Philosophical Society* in 1946. He mentioned
that, if Earth’s core had already formed, debris from a giant impact-explosion could have been metal-poor silicates, fitting the Moon’s low density. The paper remained obscure, however, and, interestingly, neither a brief Web-based biography from the American Geophysical Union nor an *Encyclopaedia Britannica* biographical sketch of Daly, both accessed in 2011, mention his work on lunar origin. A 2.5-page biographical sketch in *GSA Today* in 2006 gives half a sentence to his lunar origin work [11].

With hindsight, it appears that the reason that Daly’s work never gained much attention is that Daly was a captive of early twentieth century ideas of planetary formation, involving large masses of hot, solar-composition gases, possibly torn out of the Sun. These ideas were convincingly overturned a few years later, in the 1950s, by workers such as Schmidt [12] in Russia, and Urey [13] in the USA, who advanced geochemical evidence that planet formation involved aggregation of innumerable small cold bodies. Thus, once Daly stated his basic idea of giant impact, the rest of his discussion was doomed by its theoretical framework. For example, he describes his view of primordial Earth as follows:

According to our basal assumptions, the terrestrial gas-ball had at the beginning . . . the following characteristics: (1) a high proportion of hydrogen and helium, in which the materials of the silicate mantle and iron core . . . were ‘diluted;’ (2) a corresponding excess of mass [relative to present-day Earth]; (3) internal temperature much higher than that due merely to self-compression of the ball . . . [in a footnote, Daly cites a 1931 paper proposing a temperature at the centre of ‘something like 50 000° C.’] . . . it seems reasonable to suppose that there was a stage where the surface layer of liquid was in floating equilibrium on iron-rich . . . material [stretching to] the center of the ball. Let us assume that the liquid layer was a few hundreds of kilometers in thickness [with an] inner spheroid of ‘supercritical gas.’ . . .

Daly provides several variants of his idea. Of interest here, he titled one of them ‘External incentive to fission’ ([8], p. 108). This is similar to the above, with the following additional postulates: ‘that . . . the earth had practically its present mass and was liquid at the surface; that the planetoid had direct motion and struck the earth at or near the equator; and that fragments were torn off directly by the visitor, along with others ejected because of an explosion . . .’. This is closer to modern views, in wanting to rely at least partly on the impactor to blow the material out of Earth. As discussed in §1d, it has some resemblance to our 1986 ‘impact trigger’ proposal [14].

Because Daly’s underlying assumptions were ruled out by the 1950s, his model went virtually unmentioned during the 1960s run-up to the Apollo landings, when the existing ‘three theories’ of lunar origin were always described as: (i) co-accretion of a sister body orbiting Earth (which seemed ruled out by the Moon’s low bulk density and lack of an iron core); (ii) capture (which might explain the lack of iron if the Moon was formed far away, but was difficult to achieve, and eventually rejected due to the Moon’s isotopic near-equality with Earth); and (iii) George Darwin’s idea of fission (which was generally dismissed as ruled out by angular momentum considerations). That situation explains why we graduate students in the 1960s never heard of Daly’s idea, and why the situation ca 1970, after the first couple of lunar landings, seemed ripe for a new theory.

(c) Historical influences leading to the giant impact idea

It is perhaps of interest to recover, in retrospect, influences that led to my own formulation of a giant impact idea.

— The key influence was discovering the Orientale impact basin while working with Gerard Kuiper on his Rectified Lunar Atlas project in 1961–1962, in the first years of my graduate assistantship. The perfectly preserved, 1000-km-scale impact structure, hitherto unrecognized on the east limb of the Moon, set me wondering what were the largest
impacts that had ever occurred on Earth or in the Earth–Moon system, what were their effects, and what other gross phenomena had been missed in the 1960s’ quest for ever greater lunar detail?

— Ralph Baldwin’s 1949 book, The Face of the Moon [15], turned the tide away from volcanism and towards impact as a way of explaining the Moon’s craters, so that by the 1960s we were primed to think in those terms. He raised questions about the systematics of the mare-filled basins.

— The size distribution of lunar craters and asteroids (subject of my dissertations in the 1960s) indicated that one cannot discuss the effects of craters of one particular size without considering the effects of craters at all the other sizes. Nature is statistically unlikely to place a few big impact basins on the Moon without placing myriads of smaller impact crater as well ... and vice versa.

— Victor Safronov’s work on accretion in the 1960s (translated into English in a somewhat obscure journal, Soviet Astronomy, was followed by his book published in English in 1972 [16]). Without much parallel in the ‘West’ (except for George Wetherill), Safronov had followed mathematically the consequences of accretion. While there was some tendency in the ‘West’ to imagine a single planet emerging from among pea-sized impactors, Safronov showed that there was competitive accretion, and that a size distribution would emerge with various other sizable secondary bodies. Today’s asteroid belt appears to be an example, in which the second largest body is about half the size of the largest body. Interestingly, in the case of Uranus, Safronov estimated that an impactor, with mass ratio of $M_{\text{impactor}}/M_{\text{Uranus}}$ approximately 0.05, tipped the rotational axis to its observed state. Hartmann & Vail [17] extended similar analysis to the obliquity and rotational distributions of other planets. Around this time, Safronov’s wife, E. Ruskol [18], was a leader in co-accretion models of lunar origin, with the Earth and Moon growing in orbit around each other; this might explain why Safronov seems not to have extended his impact ideas to lunar origin.

(d) Recollections of the Kona conference

The 1984 conference in Kona, HI, USA, on the Origin of the Moon is often described as a paradigm-shifting, pivotal meeting in the history of planetary science. The concept of the meeting was strong: 10 years after the end of the Apollo programme, which was supposed to clarify the origin of the Moon and planets, planetary scientists still had no clear idea of where the Moon came from! In the run-up to the Kona conference as worldwide lunar researchers wrestled with what to say, I attempted to interest the Tucson planetary science community in the idea that modern accretion theory ensured a certain probability of catastrophic collisions, and that we should try to analyse the characteristics of such an event. Interesting resistance arose, however, because geologists had all been raised with the Hutton/Lyell uniformitarian doctrine. A giant impact smacked of long-ago-rejected catastrophism, and seemed to violate Occam’s razor by invoking an ad hoc collision. My friend and colleague Mike Drake (later head of the Lunar and Planetary Laboratory at the University of Arizona), and others, counselled me in that vein: as scientists, we needed to exhaust all non-catastrophist processes first, before considering giant impacts. The Tucson group thus submitted a paper focused on co-accretion [19]. (I participated in the discussions and joined as last author—just in case.) I submitted my own paper about giant impact [14], as did several other researchers, such as Benz et al. [20], Kipp & Melosh [21] and others, as published in the conference volume [22]. My conference presentation attacked the anti-catastrophism argument with the title ‘Stochastic ≠ ad hoc’. The idea was that contemporary accretion theory implied a significant probability of catastrophic impact occurring among the larger planetesimals, and that ‘Some classes of influential events in Solar System history are class-predictable but not event-predictable: i.e. we believe the class of events occurred, but we cannot determine times and magnitudes of individual events. These events are stochastic, but not
ad hoc’ ([14], p. 586). In retrospect, this meta-argument—that catastrophic impacts may have been a normal part of Solar System history—seemed to help clear the way for more specific giant impact concepts. ‘Stochastic ≠ ad hoc’ became the title of a subsection of [14].

The interest in some form of giant impact hypothesis was augmented by a now famous review which was entrusted to meteoriticist John Wood at the end of the meeting. In my recollection of his talk, he indicated that giant impact had emerged with the highest ranking. Wood’s published version of his ‘report card’ [23] gave no overall rankings, but assigned performance grades of A (best) to F (fail) on eight constraints (lunar mass, angular momentum, volatile element depletion, Fe depletion, O isotopes, magma ocean, physical plausibility and a lower ranked constraint involving similarity of mantle trace element patterns). Wood ranked five theories: the three classics of capture, co-accretion and fission, plus giant impact (he called it ‘collisional ejection’), and an idea he and H. E. Mitler, as well as J. V. Smith, had proposed in 1974 [24,25]. In the latter idea, which follows Ōpik [26], a body passes Earth at parabolic velocity within the Roche limit and disrupts in such a way that fragments of the outer rocky mantle on the side towards Earth are captured into orbit, but the rest of the body with its iron core speeds on by in heliocentric orbit. Here, I have reduced Wood’s rankings to a ‘grade point average’, using 1/2 weight for the mantle trace element constraint, whose value Wood said he doubted.

<table>
<thead>
<tr>
<th>hypothesis</th>
<th>John Wood ‘grade point average’ after 1984 Kona conference [23]</th>
</tr>
</thead>
<tbody>
<tr>
<td>giant impact</td>
<td>1.9</td>
</tr>
<tr>
<td>disintegrative capture</td>
<td>2.4</td>
</tr>
<tr>
<td>fission</td>
<td>2.7</td>
</tr>
<tr>
<td>co-accretion</td>
<td>3.1</td>
</tr>
<tr>
<td>capture</td>
<td>3.3</td>
</tr>
</tbody>
</table>

Interestingly, Wood’s rankings probably reflect the dominant views today (and during the 2013 Royal Society Discussion Meeting), except that ‘disintegrative capture’ has fallen by the wayside. The good reception of the giant impact hypothesis, helped by final acceptance that giant impact did not break a catastrophist taboo, encouraged a new era of more intensive modelling of giant impacts.

2. Comments on the current situation

(a) The strange case of the evidence that reversed polarity

In the early era of the giant impact hypothesis [1–3], Clayton et al. [27,28] were just discovering that objects from different parts of the Solar System have different oxygen isotope ratios, but that the Moon and Earth had virtually identical O isotope ratios, within error bars [29]. As this news spread in the mid-1970s, this fact was seen as strong support for the giant impact model, in which the Moon formed from Earth-like material, not distant material with ‘alien isotopes’.

In recent years, new data began to be reported on other elemental isotope ratios. Once again, they were virtually identical between Earth and Moon, relative to error bars, and relative to more distant bodies. One might have assumed that this would be cited as further support that lunar material was related to Earth material and that giant impact was again supported in a first-order sort of way. Instead, the evidence that had once supported the theory seemed to have ‘reversed polarity’, and now was cited against the theory. As Melosh [30] put it while proposing an ‘isotopic crisis’, ‘Unless the isotopic compositions of the proto-Earth and projectile were nearly identical by some fortuitous coincidence, there should be detectable differences between the isotopic composition of the present Earth and Moon . . .’.
The rationale for this ‘polarity reversal’ involves two steps. First, numerical modelling of giant impacts indicates that substantial fractions of the Moon’s composition come from the impactor \[31,32\]. Second, the impactor is assumed to have a significantly different isotopic composition from proto-Earth, based on dynamical models such as the Grand Tack and/or Nice models \[33\], in which distant objects with ‘alien isotope ratios’ were scattered into the Earth-forming zone, possibly from as far away as the outer Solar System (OSS), or from within the inner Solar System (ISS) \[34\]. Interestingly, this second step conflicts with the original hypothesis \[1,2\], in which the impactor was assumed to be a ‘second largest’ body in the local terrestrial zone, hence with more or less terrestrial composition. Thus, at the 2013 Royal Society Discussion Meeting, the canonical view was that, if an impact happened, the impactor had ‘alien’ isotopes and its material would make up a significant fraction of the Moon, so the Moon would necessarily have measurably different isotopic composition from Earth. It seems fair to comment that the isotopic evidence for lunar material coming from relatively Earth-like material is directly empirical, while the idea that the impactor must have had ‘alien’ isotope ratios is based on theoretical modelling.

A new twist came in 2014, when new O isotope observations by Herwartz et al. \[35\] indicated small differences between the Earth, Moon and enstatite meteorites, outside the error bars. They concluded (p. 1146) that their result ‘supports the giant impact hypothesis of Moon formation’. This may reduce the urgency of the isotope crisis. From the offset direction in O isotopes, they concluded that the impactor could not be carbonaceous, a result important in excluding an impactor from the OSS. Rather, they concluded (p. 1148) that the impactor ‘formed from the same large non-carbonaceous chondrite reservoir as Earth, Mars, ordinary chondrites, enstatite chondrites (ECs) and other non-carbonaceous chondrites and achondrites’. This view is more or less consistent with the original Hartmann/Davis \[1,2\] suggestion that the impactor was one of the largest planetesimals in the Earth-forming neighbourhood.

(b) Epistemology: what constraint exist on the belief that the putative giant impactor had ‘alien isotopes’?

Among many authors, specific arguments against Earth-like isotopic composition in the impactor have been vague. According to Melosh’s \[30\] phrase, it would have been ‘some fortuitous coincidence’. Armytage et al. \[36\], in presenting Si isotope data, stated without explanation that ‘the likelihood of such a large late impactor having the same $\Delta^{17}O$ as the proto-Earth is very low’. Zhang et al. \([37\), p. 253\], in presenting Ti isotope data, stated that, ‘The commonly accepted view … is that … radial mixing of material from different regions of the disk took place and there is \textit{a priori} no reason to expect that Theia [their name for the impactor] should have the same isotopic composition as the proto-Earth’ (they cited Chambers & Wetherill \[38\]). Meier \[39\], commenting on Zhang et al., stated without reference that ‘Theia—like all other known Solar System bodies—probably was geochemically distinct from the Earth’s mantle’.

These statements were derived ultimately from dynamical modelling. There are two classes of models, not always clearly separated. First, models such as the Grand Tack and Nice model adopt migration of the giant planets, but the degree of chaotic mixing of planetesimals is constrained. As early as 1982 \[40,41\], spectra and albedo data indicated a rough zonal correlation of the larger asteroids’ composition with the semi-major axis. Virtually all observable interplanetary bodies beyond the outer main asteroid belt have very low (a few per cent) albedos and spectral properties consistent with black, carbonaceous-type meteorites. These properties are rare in the inner main belt, at lower semi-major axes, where spectra suggest non-carbonaceous meteorite classes. As shown in figure 1, such data, when linked with asteroidal meteorite isotope properties, indicate that OSS bodies and most larger asteroid main belt bodies have isotopic properties radically different from Earth’s. Thus, any mixing resulting from migration of giant planets was insufficient to destroy the gross zonal structure among the larger interplanetary bodies. Herwartz
et al. [35] presented isotopic evidence that the impactor was not carbonaceous; they rule out an OSS impactor and argue, in effect, for an object formed in a region ranging from the inner asteroid belt (ordinary chondrites and ECs) to Earth’s zone.

The second class of earlier arguments has allowed that the impactor might have come from the ISS, but assumed a linear gradient in composition outwards from the sun, based on the difference in isotope ratios observed between Martian and terrestrial rocks. Thus, Pahlevan & Stevenson [43] pointed out in 2007 that, although the Mars–Earth difference in O isotopes was much less than the difference between OSS bodies and Earth, it was still approximately 50 times the Moon–Earth difference, so that, if O isotopes varied linearly with solar distance, the impactor would have had to form within approximately 0.02 AU of the semi-major axis of Earth. They asserted that this is virtually impossible because of the dynamical model-based arguments [34] for mixing of objects from various solar distances. Note that, according to this logic, the more nearly Earth and Moon have the same isotopic composition, the less likely a giant impact could have formed the Moon. Canup [31] thus uses such rationales ‘to estimate that the average deviation of a large impactor’s composition from that of the final planet was about half the observed compositional difference between Earth and Mars’, citing [43,44]. She then adopted a ‘Mars-like composition for our impactor . . .’, noting a few lines later that an impactor with ‘substantially more similar [composition] to Earth . . . would relax the oxygen constraint . . .’. The measurements by Herwartz et al. [35] appear to have moved the discussion empirically in that direction.

In summary, the arguments for a giant impactor having ‘alien’ isotopes are primarily based on theoretical dynamical models.

(c) Is there empirical evidence of distant early impactors with alien isotopes?

On the pro side:

— The variety of low-albedo, presumably carbonaceous retrograde and prograde captured outer satellites, including those of Mars, plus frequent carbonaceous chondrite clasts
inside other meteorite breccias, suggests scattering of OSS or outer belt carbonaceous objects into occasional satellite capture orbits and asteroid collisions, at least as far inwards as Mars [45,46].

On the con side:

— First-order zonal structure among the larger asteroids of the main belt, with different dominant spectral–taxonomic types at different solar distances [40,41], together with evidence for different isotopic ratios at different solar distances (figure 1), argue against a strong, chaotic mixing of material from different parts of the Solar System.

— The Nice model was once invoked to imply explosive scattering of OSS carbonaceous planetesimals at 3.9 Ga. The model itself did not predict any date for the scattering, but if 3.9 Ga was assumed, then it was said to confirm a suspected bombardment of the ISS, which supposedly created virtually all the observed lunar basins at that time, and the observed spike in Apollo sample impact melts at 3.85–4.0 Ga. However, subsequent study of lunar meteorites shows no spike in impacts at 3.9 Ga [47,48], nor does such a spike show up in the asteroid belt [49,50]. The Apollo spike appears associated mostly with Imbrium ejecta. These data indicate that no terminal cataclysm of ISS bombardment by OSS objects happened at 3.9 Ga, as once attributed to the Nice model. At the very least, this warns us to be cautious in assigning first-order properties of the Solar System to effects proposed in still-evolving dynamical models.

— No substantial late veneer of ‘alien’-isotope-rich material is found on the Moon. A scattering of impactors from the OSS or ISS would presumably have tapped into an entire size distribution of bodies. Thus, it is unlikely to have suddenly ended during the few-millennia accretion of the Moon after a giant impact. Continuing addition of ‘alien’ isotopes on Earth would be subducted and mixed with the mantle, but on the Moon they would create a distinctly ‘alien’ late veneer composition in the magma ocean or new crust, which was never subducted. Thus, the absence of a late ‘alien’ veneer on the Moon argues against any primordial bombardment by foreign impactors.

— Assertions that any impactor must have had strongly ‘alien isotopes’ virtually guarantee that a traditional giant impact model cannot produce the Moon with precisely Earth-like isotopes, as Melosh pointed out [30]. Such paradigms would have been more compelling if no known candidate impactors existed with Earth–Moon-like isotopic compositions. However, such a class of meteorites exists. As can be glimpsed in figure 1, ECs are considerably more isotopically Earth-like than Mars. Herwartz et al. [35], after demonstrating a small O isotope separation between Earth and Moon, find a larger separation in the same direction between Earth and the EC parent bodies. Of course, other ISS parent bodies may have existed, which are no longer sampled in our meteorite collections. Similarly, osmium isotope data suggest that impactors involved in ancient lunar upland breccias (pre-dating 4 Ga) ‘possibly represent a type of primitive material not currently delivered to Earth . . . .’ [51].

— If distant (OSS?) meteorites were more dominant in early Solar System history than in recent time, we might expect greater variety of meteorite types in the early samples than in recent samples. Joy et al., in 2013, however, suggest the opposite, that meteorite fragments from impacts more than 3.4 Ga are ‘not nearly as diverse as those found in younger . . . regolith breccias and soils . . . .’ [52]. Ancient regolith breccias, during the declining flux from 3.8 to 3.4 Ga, ‘point to chondritic precursors . . . from planetesimals that are sparsely or not at all found in the current meteorite collection . . . and the flux of comets was small’ [52]. (This is the type of impactor proposed from the most recent O isotope data [35].) Interestingly, the chemistry in lunar mare regolith soils (mostly less than 3.4 Ga) is carbonaceous [53]. Taken together, these data imply that the younger, post-3.4 Ga soils have a larger cometary, OSS component than the pre-3.4 Ga soils.
— To repeat an aspect of the original ‘isotope crisis’ argument, isotopic data from the Moon appear to rule out an OSS, giant proto-Earth impactor being involved in the Moon’s formation.

To summarize: (i) the initial arguments for impactors with alien isotopes were based on theoretical/dynamical models; (ii) empirical evidence favours some limited degree of early scattering of carbonaceous, low-albedo bodies from the OSS inwards at least as far as Mars, forming captured satellites and carbonaceous clasts in meteorite breccias; (iii) empirical evidence argues against a giant Earth-impactor from the OSS; and (iv) the Moon may have been formed by giant impact between proto-Earth and an object similar to the EC parent body, but perhaps no longer represented in the modern meteorite inventory.

(d) Where do we go from here?

Here, we consider a number of loose ends still on the table.

Our emphasis on ECs does not mean that Earth formed from such objects. Fitoussi & Bourdon [54] argued against a purely ‘EC Earth’. They suggested that Earth formed from LL, CI and CO chondrites, with a ‘maximum of 15% ECs’, fitting the above ideas of a mix of modern taxonomic types. Instead, my suggestion would be that ECs represent a ‘fringe’ population (both chemically and spatially) of the planetesimals that were forming Earth. Wasson & Wetherill ([55], p. 936) suggest they may have formed at less than 1.0 AU. The idea of enstatites as fringe objects is supported by Herwartz et al. [35] and also by early dynamical studies suggesting that a small fraction of objects spatially on the fringe of the Earth-accreting zone could eventually have been scattered into the innermost fringe of the asteroid belt [56]—exactly where we find the ‘E-belt’ that is spectroscopically suspected of containing enstatite meteorites.

Savage & Moynier [57] concluded from Si isotope evidence that EH and EL ECs represent at least two parent bodies. Further, they note that their ‘data imply that silicates condensing from a nebula gas over a wide range of ..., presumably, heliocentric distances, should have very similar Si isotope compositions’. Thus, we have four ancient bodies (Earth, Moon and two enstatite parent bodies) that ended up with approximately the same Si isotopic composition. Significant to our discussion, Savage and Moynier inferred that ‘material which condensed in the region of the solar nebula where the E-chondrites formed could still be an important component’ in terrestrial composition, hence presumably terrestrial impactors.

Some scenarios that remain on the table invoke a late impactor. Addressing this, Belbruno and Gott in 2005 [58] noted that ‘isotopic evidence suggests that [the impactor] came from 1 AU’, and suggested that it formed at a Lagrangian point, preserving it for some time before it leaked out, allowing a late impact by a large 1 AU object. Any planetesimal orbiting near 1 AU stands some chance of getting caught up in co-orbital phenomena or Lagrangian point capture, because of low relative approach velocities. Kortenkamp et al. [59] have begun further dynamical study of this process, which could favour a large, low-velocity impactor such as studied by Canup [31].

Another phenomenon might have affected all of the above scenarios. As outlined by Morbidelli [33], the Grand Tack model may have temporarily concentrated and dynamically excited the Earth–Venus–Mercury-zone planetesimals in a zone from about 0.6 to 1.0 AU. This may have produced a collisional history that tended to homogenize all these objects into more Earth-like compositions. The planets in the Earth–Venus–Mercury zone, along with the Moon-forming Earth-impactor, would thus (relative to earlier theories) have formed from bodies with more Earth-like isotopic composition. A number of speakers at the Royal Society Discussion Meeting thus urged that the most desirable empirical space programme observation would be to learn whether Venus has the same isotopic composition as Earth.

A serious aspect of the original crisis is still left, namely involving the Hf-W and Si isotope systems. As discussed at the 2013 Royal Society meeting, Hf-W and Si isotope systematics reflect the history of core–mantle evolution, and it seems unlikely for Earth and an Earth-impactor
of different size to have similar histories that would produce identical Hf-W and Si isotope systematics (cf. Touboul et al. [60]). This discussion may need to be ‘reset’, in view of [35].

At this point, it is interesting to review briefly the three Apollo-era hypotheses. The co-accretion hypothesis did not work because it could not explain a body growing next to Earth without an iron core. The capture hypothesis seems ruled out today by the inconsistencies of Earth-like isotopes arriving in an ‘alien’ object. That leaves the fission process as the only one of the three that seems to have gained from the ‘isotopic crisis’. In my paper for the Kona conference volume [14], I was concerned that the term ‘giant impact model’ might seem too simplistic, or too restrictive in terms of implied processes, and I proposed the name ‘impact-trigger hypothesis’ to allow for a wider range of processes than a one-step giant impact. The phrase is still occasionally encountered [61], but never caught on; it may be more attractive now. I visualized a triangular phase-space of impact models, with three extremes being (i) vertical impact, in which crater-like mass ejection is the main effect; (ii) near-tangential prograde impact, perhaps scattering more debris into orbit because of forward momentum and Earth’s rotation; and (iii) a near-tangential prograde impact involving ‘enough angular momentum to make Earth rotationally unstable’. Points within that triangular ‘impact phase-space’ represent various combinations of these physical effects. Some aspects of option (iii) are invoked by Ćuk & Stewart [32] and by Canup [31]. Such approaches may still be useful in fine-tuning the impact models to fit newly refined isotope ratios, reported in [35].

Impact-triggered scenarios, if physically feasible in terms of spin-up to eject terrestrial mantle material, could explain all the isotopic data, including Hf-W and Si. If the impactor was already Earth-like in isotopes, most isotope systems such as O, Zn, Ti and Cr would automatically be as observed today, and the ‘re-differentiation’ of the new Earth, while the cores of Earth and impactor merge, would produce relatively uniform Hf-W and Si systematics between the new Earth and the newly forming Moon. Much of the work on reducing high angular momentum is limited to tidal forces and related gravitational effects. However, early planets were not closed systems. As most planets apparently did not have fission-produced moons, and since the system was not sufficiently closed to require conservation of angular momentum, we can assume that later accretion of prograde and retrograde impactors had a net effect of helping to reduce later net angular momentum in the system. If the giant impactor was initially more Earth-like than in the recent ‘isotope crisis’ scenarios, as in [35], then it may be useful to consider angular momentum reduction by later impacts.

3. Conclusion

— Models of the primordial accretion and evolution of giant impactors are still evolving, and therefore we need closer examination of the assertion that early Earth impactors must come from regions with dramatically ‘alien’ isotopes, and that impactors with Earth-like isotopes would be a ‘fortuitous coincidence’.
— Insistence that either large or small late impactors were dominated by various ‘alien’ isotope systematics, based on the theoretical dynamical models, appears inconsistent with the Moon having isotope ratios virtually identical to those on Earth and also with the Moon’s lack of an upland veneer of regolith with ‘alien’ isotope properties.
— Empirical evidence, including new work in 2014 ([35], perhaps [52]), indicates that the putative giant impactor was not carbonaceous and did not come from the OSS.
— Recent work on O isotopes, showing slight offsets between Earth and Moon [35], may negate much of the ‘isotope crisis’.
— The interplay between a giant impact of ISS planetesimals, spin-up of Earth, post-impact core/mantle ‘re-differentiation’ and fission-like spin-off of ‘re-differentiated’ materials may produce models that can explain the observed W and Si isotopic properties of Earth and Moon.
— Capture of an early large planetesimal into the L4 or L5 point may allow the possibility of a late large impactor, as suggested in some giant impact models.
References


